ENERGY MODELING AND ANALYSIS IN HETEROGENEOUS CELLULAR SYSTEMS

A Dissertation
Presented to
The Academic Faculty

By

Elías Chavarría Reyes

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
in
Electrical and Computer Engineering

School of Electrical and Computer Engineering
Georgia Institute of Technology
December 2014

ENERGY MODELING AND ANALYSIS IN HETEROGENEOUS CELLULAR SYSTEMS

Approved by:

Dr. Ian F. Akyildiz, Advisor
Ken Byers Chair Professor in Telecommunications, School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Xiaoli Ma
Professor, School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Mary Ann Weitnauer
Professor, School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Geoffrey Ye Li
Professor, School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Mostafa H. Ammar
Regents' Professor, School of Computer Science
Georgia Institute of Technology

Date Approved: November 10, 2014

To my mother, father, and sister for their endless love and support.

ACKNOWLEDGMENTS

I would like to express my profound gratitude to my advisor, Dr. Ian F. Akyildiz, for his invaluable guidance and support throughout this journey. He graciously opened the doors of the Broadband Wireless Networking (BWN) Lab to me not only as a researcher, but also as a member of a wonderful family. His unbounded energy, like a perpetual flame, always ignited my passion to work harder towards my goals and my dreams. His anecdotes taught me that the road to success has a lot of ups and downs and that it is important is to celebrate the triumphs and to never give up. As a father does with his son, he genuinely shared his immense knowledge, experience, and support so that I would become the best I can be.

I would like to extend my gratitude to the professors and administrative staff of the School of Electrical and Computer Engineering at the Georgia Institute of Technology. In particular, thanks to Dr. Mary Ann Weitnauer, Dr. Xiaoli Ma, Dr. Geoffrey Li, and Dr. Mostafa Ammar, who shared their valuable time and insights as members of my Ph.D. Defense Committee.

I would also like to thank the former and current members of the BWN Lab. As siblings do, they provided me their selfless advice and support throughout these years. With them, I lived uncountable moments, both inside and outside the lab, which I take with me as unforgettable memories. It is nearly impossible to mention every member of this family; however, I give special thanks to David and Ravi, who have walked with me, side by side, throughout this journey. To all the BWN Lab members, thank you for every instant that we shared together.

Words cannot capture my gratitude towards my family; particularly, to my mother, my father, and my sister. Thank you for always believing in me, supporting me, and loving me unconditionally. Without you, this achievement would not have been possible. Last but not least, I would like to thank Galina, Valeriy, and Tatyana for their immense support during this journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Cellular network technologies have traditionally evolved to meet the ever-increasing need for capacity and coverage. As a result, several new technologies and a significant focus on small cells and heterogeneous networks have been introduced. These deployments are characterized by a large number of base stations of different types. However, the energy consumed by these base stations not only represents a significant amount of operational expenses, but is also mostly wasted due to a low energy efficiency. The traditional approach to addressing such inefficiency has been to improve the hardware components. Alternatively, better network planning and deployment strategies, adapting the active periods of the base stations according to the traffic demands, and improving the energy communication techniques have shown promising results in achieving energy consumption reduction.

The objective of this thesis is to model and analyze the energy consumption in heterogeneous cellular systems and develop techniques to minimize it. First, the energy consumption is modeled and analyzed for multi-layered heterogeneous wireless systems. This work encompasses the characterization of all the energy consumed at the base stations. Then, a novel on-off and cell-association scheme is proposed to minimize the overall network energy consumption while satisfying the spatially- and temporally-varying traffic demands. Second, we exploit the use of multi-stream carrier aggregation not only to improve the energy efficiency, but also to balance it with the conflicting objective of capacity maximization. Third, we analyze the performance of discontinuous reception methods for energy savings within the user equipments. Then, for scenarios that support carrier aggregation, we develop a cross-carrier-aware technique that further enhances such savings with minimum impact on the packet delay. Fourth, the use of small cells as an energy-saving tool and its limitations are analyzed and modeled in OPNET, a high-fidelity simulation and development platform. To bypass such limitations, a novel small cell solution is proposed, modeled, and analyzed in OPNET and then compared against its existing alternative.

CHAPTER 1

INTRODUCTION

The evolution of cellular network technologies has traditionally been driven by the ever-increasing need for capacity and coverage. The current forecasts predict that by 2019, high-speed coverage will reach over 90% of the world's population, and global traffic in mobile networks will rise with a compound annual growth rate (CAGR) of 45%, reaching a ten-fold increase since 2013 [1]. Foreseeing this exponential growth, the Third Generation Partnership Project (3GPP) completed a study in 2008 on the Long Term Evolution (LTE) for the GSM/UMTS[1] standards [2]. However, to fulfill the requirements established by the International Telecommunication Union (ITU) for IMT-Advanced[2] [3], 3GPP engaged in further enhancements of LTE, collectively known as LTE-Advanced (or LTE-A) [4]. Among the main performance objectives of LTE-A were achieving peak data rates of 1 Gbps in the downlink/500 Mbps in the uplink and maximizing the cell edge throughput. To reach these objectives, LTE-A incorporated four core technologies in its Release 10 (Rel-10) [5]: carrier aggregation, enhanced multiple input multiple output (MIMO), relay nodes, and cooperative multipoint transmission. In addition, a significant focus has been placed on exploiting more than ever before the use of small cells with heterogeneous networks in LTE-A.

1.1 Small Cells and Heterogeneous Networks

Historically, the use of small cells has been the most efficient way of increasing the capacity of cellular networks in any given area. Compared to other approaches, such as the use of more spectrum, better frequency division, and better modulation techniques, which account for capacity gains of 25, 5, and 5-fold, respectively, small cells have contributed to capacity

[1] GSM stands for global system for mobile communications. UMTS stands for universal mobile telecommunications system.

[2] IMT-Advanced stands for international mobile telecommunications-advanced.

gains of 1600-fold [6]. Furthermore, the benefits of utilizing small cells, unlike those of the other approaches, are not directly limited by Shannon's Law.

Motivated by these factors, operators initially embraced the concept of small cells by creating smaller-scale, lower-power, and lower-cost versions of the traditional macrocell base station (BS). Several terms have been used to describe these small cells, according to their coverage area: *microcell BS*, for short-range outdoor coverage; *picocell BS*, for public indoor areas or enterprise environments; and *femtocell BS*, for indoor home and office environments. Femtocells have attracted the most attention since more than 60% of all voice traffic and 70% of all data traffic originate from indoor environments [7]. Today, micro-, pico-, and femtocells are collectively referred to as *small cells* and include other technologies (e.g., Wi-Fi, relays) under their scope.

The small cells are typically deployed together with a macrocell, as shown in Figure 1. In general, they are located in areas with a high demand for an increased capacity - *the hotspots* - that the macrocell alone is not able to serve. Once a small cell is installed, it provides better service to the users near it and, most importantly, reduces the load of the macrocell, allowing the latter to better serve the rest of the users. This shift of users from the macrocell to a small cell is commonly referred to as *traffic offloading*. Since many different types of small cells may co-exist with and within a single macrocell, the term *Heterogeneous Network* (HetNet) is commonly used to describe these scenarios.

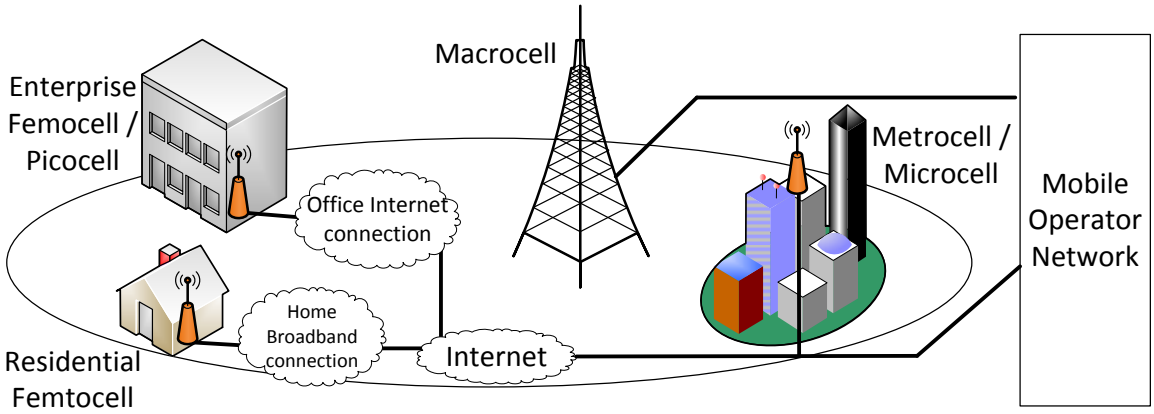


Figure 1: Heterogeneous network.

The benefits of HetNets and their presence in the future of cellular systems are undeniable [8]. Compared to macrocells, HetNets provide a more cost-effective approach to satisfy increased capacity requirements, allow traffic offloading, and are easier to deploy. However, in HetNets, many more BSs are installed in any given geographical area [9]. Each BS constantly operates at its maximum power, regardless of the amount of traffic handled, which immediately raises the concern about the environmental impact and operational expenses (OPEX) incurred by operators, particularly in terms of the energy consumption.

1.2 Energy Consumption in Heterogeneous Networks

From an economic point of view, the energy consumption is a key concern for operators, as the energy costs constitute from 7% to 20% of the entire OPEX. The network represents nearly 75% of this amount. Inside the network, close to 70% of the energy is used by the radio access network (RAN) [10], with BSs consuming the most.

From an environmental point of view, the impact of the information and communications technology (ICT) sector is also significant. Today, the ICT carbon footprint is comparable to that of the global aviation industry [11], or the one of 50 million cars. By 2020, this footprint is projected to grow at a rate of 3.8%, reaching 1.27 $GtCO_2e$[3] and contributing to 2.3% of the global green house gas (GHG) emissions [12].

Even if the operators were to disregard the environmental impact and consider the energy component of the OPEX an unavoidable cost of running the network, the energy consumption cannot be ignored. The main reason is that, with the existing technologies, the traffic is predicted to grow exponentially, while the energy efficiency shows a significantly lower, linear growth. Failing to address this mismatch will unequivocally create an unsustainable and exponentially increasing gap between the traffic growth and the energy efficiency [13].

[3] Giga tonnes of CO_2 equivalent.

Given the aforementioned reasons, both industry and academia have engaged in addressing the energy efficiency in cellular networks, also called green cellular networks, through several large-scale projects and consortia [14][15][16]. The common agreements across all these initiatives are that (i) for operators, the major energy inefficiencies and wastage occur at the RAN, and (ii) for mobile devices, the energy requirements are increasing due to the new services that demand higher data rates and quality of service (QoS).

In addition to improving the energy efficiency of the hardware components at the RAN and the user equipment (UE)[4], several approaches can be followed to reduce the energy consumption. At the RAN, these approaches include:

- **Network Planning and Deployment**. With an accurate assessment of the expected amount of traffic to be demanded by the users in a given geographic area, operators can deploy a HetNet composed of small cells to satisfy the capacity needs in hotspots and macrocells to satisfy the coverage requirement for the rest of the area of interest. Compared to a deployment with just macrocells, a HetNet can reduce the energy consumption by up to 60% [17]. Nevertheless, selecting the optimal density and location of the BSs is still an open challenge.

- **BS On-Off Schemes**. Because of the high energy consumption at the BSs and the poor energy efficiency of their components (e.g., the power amplifier (PA)), an effective way to reduce the energy consumption at the RAN is to turn the BSs off. In HetNets, this action is possible since more than one BS may be providing coverage in a particular area. The BSs must be turned off in accordance with the spatial and temporal traffic load variations that occur across different BSs throughout the day [18]. Compared to an always-on static network, a traffic-adaptive one can generate energy savings of more than 20% [19]. However, finding an optimal on-off and cell-association policy that applies to a generic m-layer HetNet remains an open issue.

[4]Following 3GPP terminology, the term "user equipment" will be used to refer to mobile devices.

- **Energy-Efficient Communication**. The same technologies introduced in LTE-A to maximize the capacity can be used to achieve energy-efficient communications. For example, adaptively switching between the multiple MIMO strategies can reduce the energy consumption by up to 50% compared to the non-adaptive approach [20][21]. Similarly, the use of relays, which have lower cost and complexity than a typical BS, allows to reduce the transmission power by a factor of five [22]. Nonetheless, several additional technologies have been introduced in LTE-Advanced, such as single- and multi-stream carrier aggregation, with the objective of improving the capacity, but without accounting for their impact on the energy consumption or exploiting them to improve the energy efficiency.

- **Small Cells for Indoor Environments**. Since a significant amount of traffic is generated from indoor environments, the use of indoor small cells is an effective method to reduce the energy consumption and encourage the traffic offloading from the macrocells. In particular, femtocells have been proposed for such environments. Nevertheless, the cross- and co-tier interference between the macrocell and the femtocells not only hinders the benefits of the latter, but also degrades the performance of the users attached to the former. Reducing the effect of both types of interference in these environments remains an open challenge.

At the UE, the concern about the energy consumption arises from the combination of two key factors. First, the UE has a very limited battery capacity. Second, the exponential traffic growth is driven by services with higher QoS requirements and data rates. Particularly, video service is expected to account for over half of the global mobile data traffic by 2019 [1]. In addition to hardware improvements, the main approach to addressing the energy consumption at the UE in LTE-A is called discontinuous reception (DRX) [23]. DRX allows the UE to turn off most of its circuitry when there are no packets to be received or transmitted. The BS, aware of the DRX state of the UE, waits for the latter to "wake up"

to send any buffered packets. Although the use of DRX has been extended to LTE-A, it currently involves simplified and inefficient approaches that fail to exploit the new features of LTE-A.

1.3 Organization of the Thesis

Achieving energy efficiency in HetNets is not a single-solution problem; rather, it requires a multi-dimensional approach that accounts for and exploits the new techniques introduced in LTE-A. This thesis encompasses such multi-dimensional approach to improve the energy efficiency not only at the RAN, but also at the UE, and is organized as follows:

- Chapter 2 analyzes the energy consumption in multi-layer HetNets. First, a BS energy consumption model is developed. It accounts for the spatio-temporal variations of the traffic demands and internal BS hardware components. Second, the problem of minimizing the energy consumption is studied and characterized. Third, an efficient algorithm is introduced to minimize the energy consumption by adjusting the on-off and cell-association policies. Then, such algorithm is evaluated in terms of its effect on the energy consumption, activity, and load across multiple layers.

- Chapter 3 investigates the use of multi-stream carrier aggregation to improve the energy efficiency in HetNets. First, the convexity of such problem is analyzed, leading to the need of a quasiconvex problem approximation. Second, from such approximation, a simple algorithm is designed to solve the energy minimization problem. Third, since operators are interested in achieving a balance between the energy minimization and capacity maximization, we develop a method to analyze this multi-objective optimization and characterize the trade-offs between the two conflicting goals.

- Chapter 4 focuses on the energy consumption at the UEs. First, we describe the operation of the LTE DRX; then, an accurate model for the LTE DRX is presented and analyzed in detail. Closed-form expressions are obtained and validated for the

achievable energy savings and the impact on the packet delay. Second, we introduce a cross-carrier-aware DRX, a novel solution for the scenarios that support carrier aggregation. We provide a detailed analysis of our proposed solution and characterize and evaluate its performance metrics against those of the classical LTE DRX.

- Chapter 5 examines the use of small cells as an energy-saving tool. First, we identify the limitations of small cells by modeling them in OPNET, a high-fidelity simulation and development platform. Second, we present our novel small cell solution, the femtorelay, capable of addressing the limitations in existing small cells. We describe in detail how the femtorelay achieves interoperability with existing cellular networks. Third, we develop a proof of concept of our femtorelay in OPNET, demonstrating the benefits of our solution in terms of not only a lower energy consumption, but also a reduced interference and an increased capacity.

- Chapter 6 provides the conclusions for this Ph.D. thesis.

CHAPTER 2

REDUCING THE ENERGY CONSUMPTION THROUGH MULTI-LAYER HETNETS

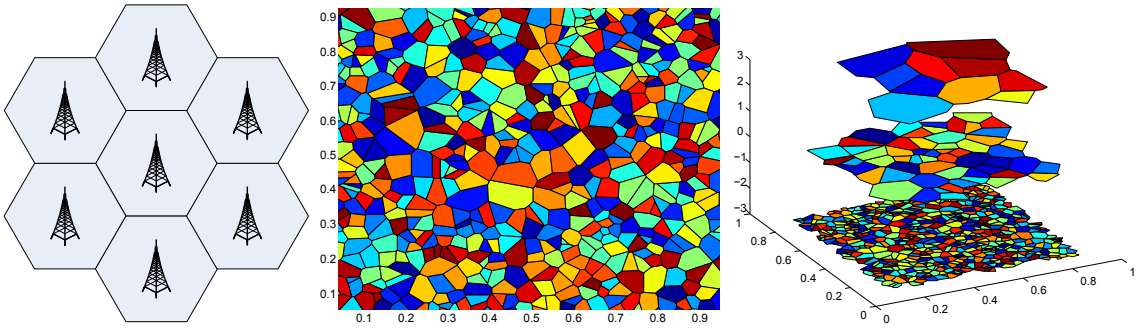
Although much research has been done to address the energy consumption in HetNets, the existing approaches have failed to capture the key factors affecting it. In this chapter, the energy consumption in HetNets is analyzed with a focus on (i) its dependence on the spatio-temporal traffic demands and internal BS hardware components and (ii) the multi-layer nature of HetNets. The problem of minimizing the energy consumption is then studied and characterized in terms of a 0-1 Knapsack-like problem. In light of the differences with the classical 0-1 Knapsack problem, an efficient algorithm is introduced to minimize the energy consumption by adjusting the cell-association and the on-off policies of the BSs. Such algorithm is shown to be applicable to the two- and m-layer HetNet cases. Performance evaluation is provided to identify the achievable energy savings of our algorithm and its effect on the energy consumption, activity, and load across multiple layers.

2.1 Motivation and Related Work

In addition to improving the hardware components of a BS [24][25], one of the more effective ways to reduce the energy consumption at the RAN is to turn the BS off. In HetNets, such action is possible since more than one BS may provide coverage in a particular area. To address the energy consumption at the RAN in HetNets, appropriate models are required for the network deployment strategies, energy consumption at the BSs, and traffic demands.

In terms of network deployment strategies, initial work [26][27][19][28][29][30][31] focused on analyzing the energy consumption in a deployment following a hexagonal grid layout [32], as shown in Figure 2a. Even though it provides useful and tractable insights regarding the general network behavior, the real deployments are far from following this

layout. Recently, a layout where BSs are deployed according to some homogeneous Poisson point process distribution, as depicted in Figure 2b, was shown to be tractable and provided a better approximation to the real network deployments [33]. This second layout was used in the analysis of the optimal BS density [34], as well as methods to adapt the active BSs and the cell-association policies [35]. Nevertheless, these two layouts assume a single layer of BSs. This is far from truth in the current systems and even more so in the future ones. Operators owning multiple frequency bands use them for multi-layer HetNet deployments, as shown in Figure 2c. The lower frequency bands are generally preferred for the BSs meant for coverage (i.e., the layer with the largest cells), while the higher frequency bands are preferred for the BSs meant for capacity (i.e., the layer with smaller cells). In contrast to the existing work, we focus on multi-layer HetNets.



(a) Hexagonal layout. (b) Homogenous Poisson point process layout. (c) Multi-layer deployment.

Figure 2: Layouts of cellular networks.

The results in the existing literature are also hindered by the BS energy consumption model used. Such model is typically a long-term function of the output power [27] [19] [28] [29] [30] [31] [36] or the overall BS traffic [35]. And even though it provides good insights, it is not readily usable to characterize the expected amount of energy that would be consumed according to the specific set of users and their traffic demands that the BS would serve. This is a serious drawback, given that the traffic demands across cellular networks can fluctuate significantly throughout the day and along BSs [37][18]. Furthermore,

since a UE in a multi-layer HetNet may be under the coverage of more than one BS [37], it is necessary to quantify the energy consumption effects of each user to determine which BS should serve a particular UE. This process is called *cell association* [38]. Its role in the energy consumption in HetNets has been previously studied [35][38], but only through simple energy models and deployment strategies. Another limitation of the existing approaches is that they generally focus on achieving simple performance thresholds, such as service outage probability, peak traffic, and minimum signal-to-interference-plus-noise ratio (SINR), among others. However, such metrics are not adequate for the analysis of the energy efficiency of a network when spatially- and temporally-varying traffic demands are considered.

In this chapter, we address the key limitations of the previous work in modeling and analyzing the energy consumption in infrastructure-based heterogeneous wireless system (IbHWS), where cellular systems are one of the cases. In particular, we model an IbHWS accounting for the multi-layer nature of HetNets and the spatio-temporal variations of the traffic demands. Then, we characterize the energy minimization problem for the two- and m-layer regular IbHWS and demonstrate its mapping to a 0-1 Knapsack-like problem. In addition, we develop and efficient algorithm to perform a joint cell-association and on-off scheme to minimize the energy consumption in a two-layer regular IbHWS. We extend this algorithm to address the m-layer case. From the simulation results, we observe energy savings of 35% from our algorithm across multiple scenarios, an inverse relationship between the network load and the energy savings, and the significant role of small cells in handling the traffic in energy-efficient deployments. More importantly, we show that the cell-association and on-off policies need to be jointly adjusted according to the actual network deployment, the energy efficiency of the BSs, and the traffic dynamics to achieve an energy-efficient network operation.

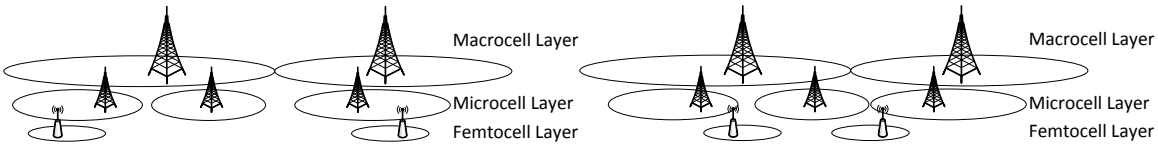
The rest of this chapter is organized as follows. We describe and analyze the network architecture, traffic model, and BS energy characterization in Sections 2.2.1, 2.2.2, and 2.2.3,

respectively. In Section 2.3, we present the energy minimization problem, its analysis for regular IbHWSs, and an energy-efficient algorithm to address the two- and m-layer cases. The performance of the algorithm is evaluated in Section 2.4. Finally, we present the conclusions in Section 2.5.

2.2 System Model

2.2.1 Network Architecture

Typically, an IbHWS is composed of multiple layers of BSs with increasing coverage area,[5] as shown in Figure 3. In Figure 3a, the coverage area of each femtocell is a subset of the coverage area of a BS in the microcell layer. Similarly, the coverage area of each microcell is a subset of the coverage area of a BS that belongs to the macrocell layer. On the other hand, in Figure 3b, the coverage area of the femtocell on the left is a subset of the coverage area not of any single microcell BS, but of a macrocell one; the coverage area of the femtocell BS on the right is not a subset of any single BS. Mathematically, these two scenarios are defined as follows:



(a) Regular IbHWS. (b) General IbHWS.

Figure 3: Infrastructure-based heterogeneous wireless system.

Definition 1: A general IbHWS is a set $\mathcal{A}$ of layers. Each layer $a \in \mathcal{A}$ contains a set $\mathcal{B}_a$ of BSs. Each BS $b \in \mathcal{B}_a$ is deployed in a coordinate $x_b \in \mathbb{R}^3$. Furthermore, there is no overlap between the effective coverage areas $\nu_{b^{(1)}} \subset \mathbb{R}^3$ and $\nu_{b^{(2)}} \subset \mathbb{R}^3$ of any two different

[5]For simplicity, we utilize the term "coverage area" to denote the 3D coverage space. Similarly, diagrams will depict 2D versions of the coverage area for the ease of understanding. Nevertheless, the analysis is done considering a 3D coverage space.

BSs $b^{(1)}$ and $b^{(2)}$ that belong to the same layer a, i.e.,

$$\nu_{b^{(1)}} \cap \nu_{b^{(2)}} = \emptyset, \qquad \forall a \in \mathcal{A}, b^{(1)} \in \mathcal{B}_a, b^{(2)} \in \mathcal{B}_a, b^{(1)} \neq b^{(2)}. \tag{1}$$

The effective coverage area represents the area in which users will prefer to connect to BS b rather than to any other BS belonging to the same layer as b.

In general, the BSs within the same layer share certain general characteristics, but may differ in their hardware and configurations. For example, the coverage area of the BSs in the femtocell layer is typically just enough for residential or small office environments. However, different femtocells may vary in their number of antennas, power amplifier energy efficiency, and other hardware characteristics. Moreover, even two femtocells with the exact same hardware may be configured to operate differently, e.g., with different transmission power.

Definition 2: A regular IbHWS is a general IbHWS for which there exists an ordering of layers $a_0, a_1, ...$ so that $\forall b^{(i)} \in a_i, \forall b^{(j)} \in a_j, j > i$:

$$\textbf{if } \nu_{b^{(i)}} \cap \nu_{b^{(j)}} \neq \emptyset \qquad \textbf{then} \qquad \nu_{b^{(i)}} \subseteq \nu_{b^{(j)}}. \tag{2}$$

Intuitively, in a regular IbHWS, any given BS coverage area is either the largest in a particular geographic segment or a subset of a larger coverage area that belongs to another BS.

2.2.2 Traffic Model

In the existing literature, a traffic demand (TD) is typically characterized by the session arrival rate and the average file size [35]. With these two values, the overall load of a BS is then estimated in terms of (a) the total number of bits served by the BS, or (b) the average bit rate, i.e., the total number of bits divided by the total time. Then, based on the value of overall load, the energy consumption of the BS is predicted. However, this approach is

highly inaccurate, mainly because it assumes a linear relationship between a particular bit rate and the corresponding power required to provide it.

If we were to transmit the same number of bits ρ at two different bit rates R_1 and R_0, where $R_1 = mR_0$ and $m > 1$, then the amounts of time t_1 and t_0 required at each rate, respectively, would be related by

$$t_0 = mt_1. \tag{3}$$

Therefore, for the energy consumption at both rates to be equal, it would be necessary that

$$P_{RX_1} = mP_{RX_0}, \tag{4}$$

where P_{RX_1} and P_{RX_0} represent the received power required to achieve each bit rate, respectively. However, utilizing Shannon's Channel Capacity theorem, it can be shown that the relationship between P_{RX_1} and P_{RX_0} is exponential rather than linear:

$$P_{RX_1} = \left[\left(1 + \frac{P_{RX_0}}{\eta}\right)^m - 1\right]\eta, \tag{5}$$

where η represents the noise plus interference. Even if the analysis is restricted to a unique bit rate, i.e., $m = 1$, the value of P_{RX} will vary with the distance between the BS and the intended receiver. In conclusion, energy consumption estimation based on the classical traffic demand modeling is inaccurate due to (a) the non-linear relationship between the power and the bit rate and (b) the distance-dependent power required to satisfy a particular bit rate.

To accurately characterize the energy consumption, we follow a different approach. We consider the network to be able to serve sessions that have different QoS requirements. These are defined in terms of bit rates, to capture the non-linear relationship between the bit rate and the power. Here, $\mathcal{Q}$ denotes the set of QoS requirements that the network supports.

We consider the total service area $\mathcal{U}$ to be divided into locations $u \subset \mathbb{R}^3$, satisfying the

following requirements:

1. Each location u is a connected set.
2. Locations are non-overlapping: $u_1 \cap u_2 = \varnothing$.
3. Each location can generate sessions of any QoS $q \in \mathcal{Q}$.
4. $\mathcal{U} = \bigcup u$.

For every BS and the locations it serves, we use a matrix $N(t)$ to characterize the number of active sessions at any time instant t corresponding to every location u and QoS q.

2.2.3 Base Station Energy Characterization

In this section, we analyze the total energy that a BS consumes to satisfy a set of traffic demands. As we later describe, this energy depends not only on the radio frequency (RF) transmission power required to serve each traffic demand, but also on the energy consumed by the internal components of the BS. Characterizing such consumption is critical, since it is the highest within the BS.

2.2.3.1 RF Energy

Here, we focus on mapping the traffic demands a BS needs to satisfy into the RF energy required by that BS. Consider a BS b that serves a set of locations $\mathcal{G} \subset \mathcal{U}$. The total RF power radiated by b is

$$P_{RF}(t) = \mathrm{tr}\left(N(t)P^T(t)\right), \tag{6}$$

where $\mathrm{tr}(A)$ denotes the trace of a matrix A. $N(t)$ and $P(t)$ are matrices whose elements $N_{\omega,\tau}(t)$ and $P_{\omega,\tau}(t)$ denote, at a time t and location $u_\omega \in \mathcal{G}$, the number of sessions with QoS q_τ and the amount of power to satisfy one session of QoS q_τ.

The total RF energy radiated by b during the j-th time interval[6] Δt_j of the day is

$$E_{RF}(\Delta t_j) = \int_{\Delta t_j} P_{RF}(t)\mathrm{d}t = \int_{\Delta t_j} \mathrm{tr}\left(N(t)P^T(t)\right)\mathrm{d}t. \tag{7}$$

[6]Each time interval is in the order of 30 minutes to 1 hour.

Assuming that $N(t)$ and $P(t)$ are uncorrelated, the expected value of this energy is

$$\mathbb{E}[E_{RF}(\Delta t_j)] = \int_{\Delta t_j} \mathrm{tr}\left(\mathbb{E}\left[N(t)\right]\mathbb{E}\left[P^T(t)\right]\right)\mathrm{d}t. \tag{8}$$

In terms of the existing wireless networks, this assumption implies that the power required to serve a user at a specific rate is uncorrelated with the number of users that are served by the BS. This assumption can hold for existing systems. For example, in systems based on OFDMA[7], such as LTE, once the subcarriers are allocated to a user, the power required to serve the user is independent of the number of users served by the same BS. In UMTS/WCDMA[8], this assumption generally does not hold because of the orthogonality factor greater than zero. However, the expression can still apply if the power is calculated assuming an expected level of intra-cell interference.

We can further simplify this expression by considering that $N(t)$, $P(t)$, or both are first-order stationary (FOS). In the case that both are FOS, we get

$$\mathbb{E}[E_{RF}(\Delta t_j)] = \Delta t_j \,\mathrm{tr}\left(\mathbb{E}\left[N\right]\mathbb{E}\left[P^T\right]\right), \tag{9}$$

where the time dependency of N and P has been removed to reflect the first-order stationarity of each random process.

To produce the amount of RF energy described by Eq. (9), the BS consumes a significant amount of energy internally, as is discussed in the next section.

2.2.3.2 Base Station Energy

We model the BS as a set of interconnected components $c \in C$. These components represent the major energy-consuming elements in a BS, such as the feeder, power amplifier (PA), baseband microprocessor (BBuP), power supply (PS), and air conditioner (A/C), as shown in Figure 4 [39]. We observe that such components consume most of the energy that the

[7]OFDMA stands for orthogonal frequency division multiple access.

[8]WCDMA stands for wideband code division multiple access.

BS receives as input, while the radiated RF energy is a relatively small percent of the total BS energy consumption.

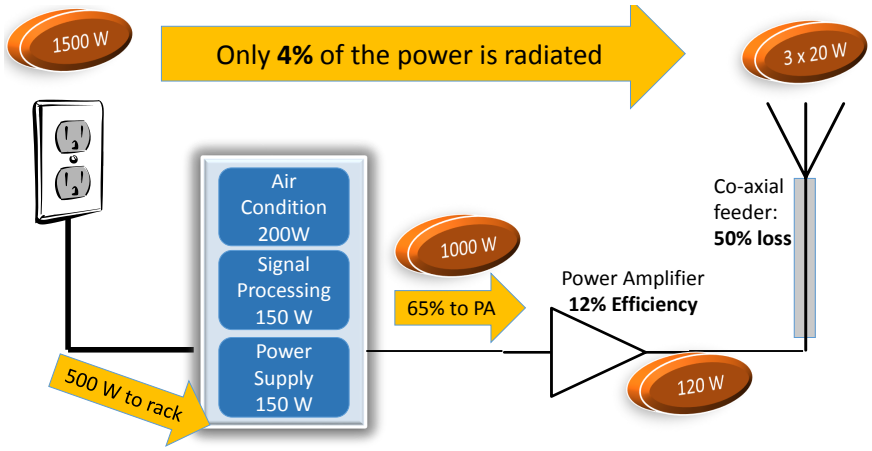


Figure 4: Major energy-consuming elements in a macrocell base station, considering three sectors with 20W of output RF power per sector.

Each component can be either *on* (with load), *on* (with no load), or *off*, and it continuously consumes at least a minimum amount of energy, whether it is *on* or *off*:

$$E_{\text{off}}(\Delta t_j) = P_{\text{off}} \Delta t_j, \tag{10}$$

$$E_{\text{on,min}}(\Delta t_j) = P_{\text{on,min}} \Delta t_j, \tag{11}$$

where P_{off} is the standby power, and $P_{\text{on,min}}$ is the power consumed with no load. For passive components,

$$P_{\text{on,min}} = P_{\text{off}} = 0, \tag{12}$$

and for non-passive components,

$$P_{\text{on,min}} \geq P_{\text{off}} \geq 0. \tag{13}$$

In addition to the minimum energy consumed in the *on* state, each component also has a dynamic energy consumption $E_{\text{on,dyn}}(\Delta t_j)$ that can follow one of two models. In the first

model, $E_{\text{on,dyn}}(\Delta t_j)$ is a function of the energy that the component needs to produce at its output:

$$E_{\text{on,dyn}}(\Delta t_j) = \int_{\Delta t_j} \frac{1}{\alpha(t)} P_{\text{out}}(t) \mathrm{d}t, \tag{14}$$

where $\alpha(t)$ is the power efficiency of the device, and $P_{out}(t)$ is the power that it needs to produce at its output. For most components, $\alpha(t)$ can be assumed to be a constant. So, taking the expected value, we obtain

$$\mathbb{E}[E_{\text{on,dyn}}(\Delta t_j)] = \frac{1}{\alpha} \mathbb{E}[E_{\text{out}}(\Delta t_j)]. \tag{15}$$

We denote the components following this model as Type A. These include the PA and the feeder.

In the second model, $E_{\text{on,dyn}}(\Delta t_j)$ is a function of the overall traffic load that it needs to process:

$$E_{\text{on,dyn}}(\Delta t_j) = \int_{\Delta t_j} (P_{\text{on,max}} - P_{\text{on,min}}) h(\text{load}(t)) \mathrm{d}t, \tag{16}$$

where $P_{\text{on,max}}$ is the maximum power, and $0 \leq\ h(\text{load}(t)) \leq 1$. By assuming that h is linear in the load and that the load is FOS, the expected value of the dynamic energy becomes

$$\mathbb{E}[E_{\text{on,dyn}}(\Delta t_j)] = \frac{\Delta t_j (P_{\text{on,max}} - P_{\text{on,min}})}{\text{load}_{\text{max}}} \mathbb{E}[\text{load}]. \tag{17}$$

$\mathbb{E}[\text{load}]$ can be expressed as $\text{tr}\left(\mathbb{E}\left[\text{N(t)R}^{\text{T}}\right]\right)$, where R is a matrix in which (i) all rows are identical, and (ii) the values in each row represent the bit rate associated with each $q \in \mathcal{Q}$. Assuming FOS for $N(t)$, we obtain

$$\mathbb{E}[E_{\text{on,dyn}}(\Delta t_j)] = \frac{\Delta t_j (P_{\text{on,max}} - P_{\text{on,min}})}{\text{load}_{\text{max}}} \text{tr}\left(\mathbb{E}[N] R^T\right). \tag{18}$$

We denote the components following this second model as Type B. An example is the microprocessor.

Remark 1: The energy consumption model of the BS is an affine function of $\mathrm{tr}\left(\mathbb{E}[N]R^T\right)$ and $\mathrm{tr}\left(\mathbb{E}[N]\,\mathbb{E}\left[P^T\right]\right)$.

Proof: In this section, all the energy models are affine functions of $\mathrm{tr}\left(\mathbb{E}[N]R^T\right)$ and $\mathrm{tr}\left(\mathbb{E}[N]\,\mathbb{E}\left[P^T\right]\right)$. Since the interconnection of any two components leads to the composition of their respective energy models, their joint energy model is also an affine function of $\mathrm{tr}\left(\mathbb{E}[N]R^T\right)$ and $\mathrm{tr}\left(\mathbb{E}[N]\,\mathbb{E}\left[P^T\right]\right)$. By induction, it follows that the energy consumption model of the BS is also an affine function of these two variables. ■

Being an affine function, the energy consumption $\hat{E}_b(\Delta t_j)$ of a BS b can be expressed in the form of

$$\hat{E}_b(\Delta t_j) = \begin{cases} \hat{E}_{\text{off}}(\Delta t_j) & \text{if } b \text{ is } \textit{off} \\ \hat{E}_{\text{on,min}}(\Delta t_j) + \hat{E}_{\text{on,dyn}}(\Delta t_j) & \text{if } b \text{ is } \textit{on} \end{cases}, \tag{19}$$

where $\hat{E}_{\text{off}}(\Delta t_j)$ represents the total energy consumption of b in the *off* state. $\hat{E}_{\text{on,min}}(\Delta t_j)$ and $\hat{E}_{\text{on,dyn}}(\Delta t_j)$ represent the total minimum and dynamic energy consumption, respectively, when b is *on*. $\hat{E}_{\text{on,dyn}}(\Delta t_j)$ can be further expressed as the linear function:

$$\hat{E}_{\text{on,dyn}}(\Delta t_j) = A \begin{bmatrix} \mathrm{tr}\left(\mathbb{E}[N]R^T\right) \\ \mathrm{tr}\left(\mathbb{E}[N]\,\mathbb{E}\left[P^T\right]\right) \end{bmatrix}, \tag{20}$$

where A is a 1x2 matrix unique to every BS. The exact expressions for $\hat{E}_{\text{off}}(\Delta t_j)$, $\hat{E}_{\text{on,min}}(\Delta t_j)$, and A depend on the specific set C of components present in a BS b and how they are interconnected. For example, even if no user is connected to a BS b, the power supply is *on* (with load) since it provides the power to all components.

Our affine model for the BS energy consumption fits the early measurements reported in [29], as part of the EARTH[9] project. However, in [29], the model was approximated from the measured data rather than explicitly derived, as done here. Furthermore, the model in [29] is limited since it only considers the RF output power dependency. In addition, it is not suitable to predict the BS energy consumption in specific traffic distributions.

[9]EARTH stands for Energy Aware Radio and neTwork tecHnologies.

Remark 2: $\mathrm{tr}\left(\mathbb{E}[N]R^T\right)$ is a linear function along the dimension of the locations u.

Proof: The i-th element of the main diagonal of the matrix $\mathbb{E}[N]R^T$ can be expressed as

$$\sum_{\tau} \mathbb{E}\left[N_{i,\tau}\right] R_{\tau,i}, \tag{21}$$

where $N_{i,\tau}$ is the number of active sessions of QoS q_τ from location u_i, and $R_{\tau,i}$ is the rate associated with QoS q_τ for a location u_i. Therefore, the i-th element of the main diagonal is uniquely defined by the values from the location u_i. It follows that the summation of the elements of the main diagonal of the matrix $\mathbb{E}[N]R^T$, i.e., the trace function, is a linear function along the dimension of the locations u. ■

Remark 3: $\mathrm{tr}\left(\mathbb{E}\left[N\right]\mathbb{E}\left[P^T\right]\right)$ is a linear function along the dimension of the locations u.

Proof: The i-th element of the main diagonal of the matrix $\mathbb{E}\left[N\right]\mathbb{E}\left[P^T\right]$ can be expressed as

$$\sum_{\tau} \mathbb{E}\left[N_{i,\tau}\right]\mathbb{E}\left[P_{\tau,i}\right], \tag{22}$$

where $P_{\tau,i}$ is the RF power required to satisfy a single session of QoS q_τ for a location u_i. Following the same argument as in Remark 1, we see that $\mathrm{tr}\left(\mathbb{E}\left[N\right]\mathbb{E}\left[P^T\right]\right)$ is a linear function along the dimension of the locations u. ■

Remark 4: $\hat{E}_{\mathrm{on,dyn}}(\Delta t_j)$ is a linear function along the dimension of the locations u.

Proof: Since $\hat{E}_{\mathrm{on,dyn}}(\Delta t_j)$ is a linear function of $\mathrm{tr}\left(\mathbb{E}[N]R^T\right)$ and $\mathrm{tr}\left(\mathbb{E}\left[N\right]\mathbb{E}\left[P^T\right]\right)$, and the two are linear functions along the dimension of the locations u, it follows that $\hat{E}_{\mathrm{on,dyn}}(\Delta t_j)$ is a linear function along the dimension of the locations u. ■

Therefore, the energy consumption of a BS b, $\hat{E}_b(\Delta t_j)$, is itself an affine function along the dimension of the locations u. This allows us, in section 2.3, to analyze the effect on the energy consumption $\hat{E}_b(\Delta t_j)$ of the traffic coming from each location that a BS can potentially serve.

2.3 Joint Cell-Association and On-Off Scheme

For the j-th time interval Δt_j, the maximum energy savings at the RAN can be achieved by solving the following optimization problem:

$$\text{minimize} \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{B}_a} \mathbb{E}\left[\hat{E}_b(\Delta t_j)\right], \tag{23}$$

subject to:

$$\sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{B}_a} \phi_b(\Delta t_j, u) = 1, \forall u \in \mathcal{U}, \tag{24a}$$

$$\sum_{u \in \Psi_b(\Delta t_j)} \sum_{q \in Q} \mathbb{E}\left[N_{u,q}\right] R_{u,q} \leq \Upsilon_b, \forall a \in \mathcal{A}, \forall b \in \mathcal{B}_a, \tag{24b}$$

where $\phi_b(\Delta t_j, u)$ is an indicator function equal to 1 when b serves u, $\Psi_b(\Delta t_j)$ is the set of users served by b, and Υ_b is the maximum load supported by b. Constraint (24a) indicates that each u is served by a single b, while constraint (24b) indicates that the total load in each BS should be below the maximum that BS can support. This problem is in general non-convex and not solvable through standard optimization methods. Nevertheless, the complexity of obtaining the solution at the j-th time interval is reduced if any of the following conditions apply:

1. For every active location u that can only be served by a single BS, that BS must be in the *on* state. For any such BS, its energy contribution to the overall RAN energy consumption is at least $\hat{E}_{\text{on,min}}(\Delta t_j)$.

2. Every BS b for which no location under its coverage requests any traffic can be switched to the *off* state. For any such BS, its energy contribution to the overall RAN energy consumption is determined by Eq. (19).

   From an analytical perspective, this condition should always be enforced. However, in a real IbHWS, turning a cell off should only be done when it is highly probable

that no user will appear in that area (e.g., an office femtocell during the night), or when another active cell is serving the same area.

An alternative way of looking at the energy minimization problem is to solve two coupled problems:

- Cell (de-)activation: Which BSs should be turned on? Which ones should be turned off?
- Cell association: To which BS should each location connect to be served?

If the answer to either one of these two problems is known, then the optimal solution for the other problem can be found. Rather than considering these two problems separately, we focus on solving them jointly for the regular IbHWS.

Remark 5: Given a regular IbHWS $\mathcal{A}$, there exists a partition Γ of the total coverage area of interest

$$\Xi = \bigcup_{\substack{b \in \mathcal{B}_a \\ a \in \mathcal{A}}} \nu_b, \tag{25}$$

such that

$$\forall \gamma \in \Gamma : \exists a \in \mathcal{A}, b \in \mathcal{B}_a \mid \gamma = \nu_b. \tag{26}$$

Proof: The partition Γ is obtained through the following steps. First, make Γ equal to a set of all possible ν_b. Then, remove any ν_b that is a subset of another element of Γ due to condition (2). Elements that were not removed must satisfy the condition that their intersection with any other element of Γ is the empty set. Therefore, the remaining elements form a partition of Ξ. ■

From the proof above, it also follows that

$$\forall a \in \mathcal{A}, b \in \mathcal{B}_a : \exists \gamma \in \Gamma \mid \nu_b \subseteq \gamma. \tag{27}$$

Therefore, any IbHWS $\mathcal{A}$ can be partitioned into mutually disjoint areas, such that the effective coverage area of each BS is a subset of some element of the partition. At the same time, all the BSs for which $\nu_b \subseteq \gamma$ for a particular $\gamma \in \Gamma$ also constitute an IbHWS where the last layer has a single BS. Figure 5 depicts this partitioning for the regular IbHWS in Figure 3a.

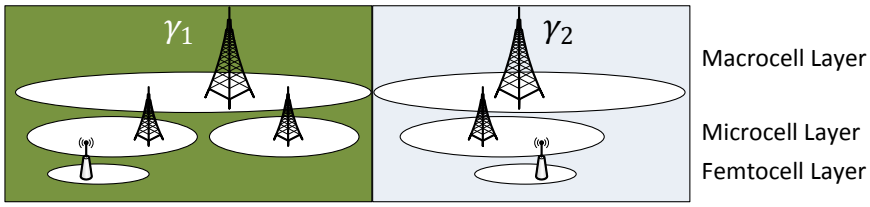


Figure 5: Partition of regular IbHWS.

Remark 6: For a regular IbHWS, the optimal solution to the energy minimization problem is achieved by finding the optimal solution for each IbHWS defined by the elements of Γ.

Proof: From Eq. (27), it follows that the on-off configuration of any BS b only affects the energy consumption of the IbHWS defined by the γ for which $\nu_b \subseteq \gamma$. Thus, the IbHWSs defined by the rest of the elements of Γ can be independently configured from the on-off state of b. ■

In conclusion, finding the optimal solution for the energy minimization problem in a regular IbHWS reduces to being able to find the optimal solution for a regular IbHWS whose last layer has a single BS. We exploit this property to develop an efficient algorithm to minimize the energy consumption for the two- and multi-layer regular IbHWS.

2.3.1 Two-Layer HetNet Analysis

A two-layer regular IbHWS is depicted in Figure 6. While simple, this type of scenario is important since it is the most commonly deployed when operators are migrating their networks from a single-layer layout to a multi-layer one. Furthermore, it allows us to establish the ideas that become the building blocks for the m-layer regular IbHWS.

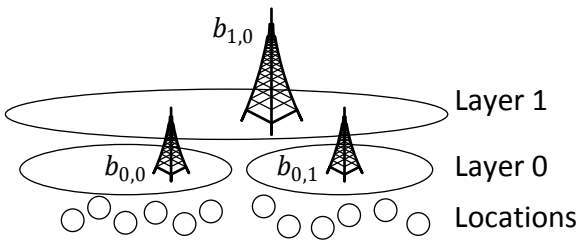


Figure 6: Two-layer regular IbHWS.

The approach to addressing the energy minimization problem at the j-th time interval for a two-layer regular IbHWS is now presented. For simplicity, we utilize $b_{i,k}$ to denote the i-th BS of layer $a_k \in \mathcal{A}$ and drop the Δt_j dependency.

First, we consider that layer a_1 is not present. Therefore, every location u must be served by a BS $b_{i,0}$, i.e., that belongs to layer a_0, and we can obtain the energy consumption of the network by applying Eq. (19) to every BS $b_{i,0}$. We denote the network energy consumption for this case by E_{N-0}. For n BSs, the complexity of computing this value is $O(n)$.

Second, we consider that layer a_1, which only has one BS $b_{0,1}$, is also present. At this point, we have a decision to make: should $b_{0,1}$ be *on* or *off*? To answer this question, we need to compare the minimum energy required by the network when $b_{0,1}$ is *off* against the minimum one required when $b_{0,1}$ is *on*. We will now find these two values.

If $b_{0,1}$ is *off*, the overall network consumption E_{N-1} becomes

$$E_{N-1} = E_{N-0} + \hat{E}_{\text{off}}(b_{0,1}), \tag{28}$$

where $\hat{E}_{\text{off}}(b_{0,1})$ corresponds to the energy consumed by $b_{0,1}$ while *off*, as defined in Eq. (19).

If $b_{0,1}$ is *on*, the network consumption must increase by at least $\hat{E}_{\text{on,min}}(b_{0,1})$, as defined in Eq. (19). Therefore, in this case,

$$E_{N-1} \geq E_{N-0} + \hat{E}_{\text{on,min}}(b_{0,1}). \tag{29}$$

When $b_{0,1}$ is *on* and all locations are served by layer a_0, there are two types of actions that

can be further executed to reduce the energy consumption E_{N-1}.

1. **Action Type 1:** Use $b_{0,1}$ to a serve a particular location u instead of layer a_0. For this action to be executable, the following condition must be satisfied:

$$\text{RemCap}(b_{0,1}) \geq \text{ReqCap}(\text{AT1}_u), \tag{30}$$

where $\text{RemCap}(b_{0,1})$ denotes the remaining capacity in $b_{0,1}$, and $\text{ReqCap}(\text{AT1}_u)$ is the required capacity to serve location u. If executable, the change in the energy consumption $\Delta E_1(u)$ of the overall network is

$$\Delta E_1(u) = \hat{E}_{\text{on,dyn}}(b_{0,1}, u) - \hat{E}_{\text{on,dyn}}(b_{i,0}, u), \tag{31}$$

where $b_{i,0}$ is the BS of layer a_0 initially serving the particular location u. However, if the location u was the only one served by $b_{i,0}$, then such BS can be turned off. Therefore, an additional change in the overall energy consumption takes place:

$$\Delta E_1(u) = \Delta E_1(u) + \hat{E}_{\text{off}}(b_{i,0}) - \hat{E}_{\text{on,min}}(b_{i,0}). \tag{32}$$

From Eq. (13), it immediately follows that

$$\hat{E}_{\text{off}}(b_{i,0}) - \hat{E}_{\text{on,min}}(b_{i,0}) \leq 0. \tag{33}$$

A value $\Delta E_1 < 0$ implies a reduction in the overall energy consumption of the network.

2. **Action Type 2:** Use $b_{0,1}$ to serve all locations u originally served by a particular BS $b_{i,0}$. For this action to be executable, the following condition must be satisfied:

$$\text{RemCap}(b_{0,1}) \geq \text{ReqCap}(\text{AT2}_{b_{i,0}}), \tag{34}$$

where ReqCap($AT2_{b_{i,0}}$) denotes the required capacity to serve all locations u originally served by $b_{i,0}$. If executable, the change in the overall energy consumption $\Delta E_2(b_{i,0})$ of the network is

$$\Delta E_2(b_{i,0}) = \hat{E}_{\text{off}}(b_{i,0}) + \sum_{u \in \Psi_{b_{i,0}}} E_{\text{on,dyn}}(b_{0,1}, u)$$
$$- \hat{E}_{\text{on,min}}(b_{i,0}) - \sum_{u \in \Psi_{b_{i,0}}} E_{\text{on,dyn}}(b_{i,0}, u), \tag{35}$$

where Ψ_b is the set of locations served by b. A value $\Delta E_2 < 0$ implies a reduction in the overall energy consumption of the network.

Remark 7: Eq. (35) reduces to Eq. (32) when $b_{i,0}$ serves a single location u.

Remark 8: The execution of Action Type 2 can also be interpreted as the application of Action Type 1 to all u originally served by $b_{i,0}$, followed by turning off that $b_{i,0}$.

For n BSs in layer a_0 and p locations, $p \geq n$, calculating the effect ΔE_1 of Action Type 1 for all locations u has an overall complexity of $O(p)$. Similarly, calculating the effect ΔE_2 of Action Type 2 for all $b_{i,0}$ also has an overall complexity of $O(p)$.

Utilizing the above formulation, we now describe how the energy saving problem is mapped to a Knapsack-like problem. Consider the following concept mapping:

- $-\Delta E_1(u)$: "Profit" from executing Action Type 1 on u.
- $-\Delta E_2(b_{i,0})$: "Profit" from executing Action Type 2 on $b_{i,0}$.
- ReqCap($AT1_u$): "Weight" of executing Action Type 1 on u.
- ReqCap($AT2_{b_{i,0}}$): "Weight" of executing Action Type 2 on $b_{i,0}$.
- Cap($b_{0,1}$): Maximum "weight" supported by BS $b_{0,1}$.

Then, finding the minimum energy consumed by the network when $b_{0,1}$ is *on* can be represented as a 0-1 Knapsack-like problem, where each action has a "weight" and provides a

"profit". The major differences from the traditional 0-1 Knapsack problem are

- The "profit" of an action may be negative. This occurs if the action causes an increase in the overall energy consumption of the network. No such action should be executed.
- The "profit" of an action may change depending on previously executed actions. For example, if Action Type 1 is executed on a location u, then the additional achievable profit of executing Action Type 2 on the serving BS $b_{i,0}$ of u will be less than the original value of $-\Delta E_2(b_{i,0})$.

For these reasons, we introduce Algorithm 1 to address the Knapsack-like problem of finding the minimum energy consumption when $b_{0,1}$ is *on*. Without loss of generality, we assume that every location u could be served by some $b_{i,0}$. If this were not the case, the algorithm would be modified so that $b_{0,1}$ served such locations before executing any other action.

Algorithm 1 Minimum E_{N-1} for $b_{0,1}$ *on*.

```
1:  Call Init-Vars                                          ▷ Calls Algorithm 2
2:  for i ← 1, |H| do
3:      h ← i-th element from H
4:      ATX ← action associated with h
5:      if RemCap(b_{0,1}) ≥ ReqCap(ATX) then
6:          Call Execute-Action                             ▷ Calls Algorithm 3
7:      else
8:          Discard ATX
9:      end if
10: end for
```

The execution of Algorithm 1 depends on the initialization performed in Algorithm 2. In the latter, Eff() denotes a function that calculates the efficiency of an action as

$$\text{Eff}() = \frac{\Delta E}{\text{ReqCap}}. \tag{36}$$

Algorithm 1 also depends on the execution of Algorithm 3. In the latter, Prev-mod(b) indicates whether one or more locations originally served by b have already been reassigned

Algorithm 2 Init-Vars.

1: $E_{N-1} = E_{N-0} + \hat{E}_{\text{on,min}}(b_{0,1})$
2: **for all** u **do**
3: Calculate $\Delta E_1(u)$
4: **if** $\Delta E_1(u) > 0$ **then**
5: Discard AT1_u
6: **else**
7: Calculate $\text{Eff}(\text{AT1}_u)$
8: **end if**
9: **end for**
10: **for all** $b_{i,0}$ **do**
11: Calculate $\Delta E_2(b_{i,0})$
12: **if** $\Delta E_2(b_{i,0}) > 0$ **then**
13: Discard $\text{AT2}_{b_{i,0}}$
14: **else**
15: Calculate $\text{Eff}(\text{AT2}_{b_{i,0}})$
16: **end if**
17: **end for**
18: List $H \leftarrow$ Jointly sort all $\text{Eff}(\text{AT1}_u)$ and $\text{Eff}(\text{AT2}_{b_{i,0}})$ in increasing order

to $b_{0,1}$, and $\Delta E_{\text{previous}}$ denotes the energy savings achieved by the previous reassignment of such locations. In Algorithm 3, lines 7-9 and 15-16 adjust the application of a new action depending on previously executed actions.

Together, these three algorithms have an overall complexity of $O(p\log(p))$, which is dominated by the sorting performed in the last step of Algorithm 2. These algorithms allow us to:

1. Find the minimum amount of energy consumed when $b_{0,1}$ is *on*, which we can compare to the energy required when $b_{0,1}$ is *off* (Eq. (28)).

2. Jointly determine which BSs should be *on* or *off* and with which BS each location should be associated.

Remark 9: The energy minimization problem reduces to the classical 0-1 Knapsack problem if

$$\hat{E}_{\text{off}}(b_{i,0}) = \hat{E}_{\text{on,min}}(b_{i,0}), \qquad \forall i. \tag{37}$$

Algorithm 3 Execute-Action.

if Type(ATX) = Type 2 **then**
  $b \leftarrow$ BS associated with ATX
  if Prev-Mod(b) = False **then**
    Apply ATX, and Update RemCap($b_{0,1}$)
    $E_{N-1} = E_{N-1} + \Delta E_2(b)$
  else
    if $\Delta E_2(b) < \Delta E_{\text{previous}}$ **then**
      Apply ATX, and Update RemCap($b_{0,1}$)
      $E_{N-1} = E_{N-1} + \Delta E_2(b) - \Delta E_{\text{previous}}$
    end if
  end if
else ▷ Type 1
  $u \leftarrow$ location associated with ATX
  $b \leftarrow$ servings BS of u
  if b is alreay *off* **then**
    Discard ATX
  else ▷ b is *on*
    Apply ATX, and Update RemCap($b_{0,1}$)
    $E_{N-1} = E_{N-1} + \Delta E_1(u)$
  end if
end if

Proof: When the previous condition holds, applying Action Type 2 to a BS b has the same effect on the overall energy consumption as the application of Action Type 1 to all the locations served by b. Therefore, we can omit Action Type 2 from the problem formulation and only consider Type 1 actions. ■

The condition established in Remark 9 applies to most BSs nowadays since they lack the capability to enter into a dormant state for energy savings. However, in the most recent studies for 4G networks, such as LTE-A, the ability to enter into a dormant state is being introduced.

2.3.2 Multi-Layer HetNet Analysis

In this section, we use the scheme from the two-layer regular IbHWS to develop a solution for the m-layer regular IbHWS. The proposed solution for the m-layer regular IbHWS is efficiently obtained from the solution of the $(m-1)$-layer regular IbHWS, making the

approach scalable.

Consider a regular IbHWS $\mathcal{A}$ of m layers:

$$\mathcal{A} = \{a_0, a_1, ..., a_{m-1}\}, \tag{38}$$

where the top layer a_{m-1} has a single BS. It is enough to analyze this case alone since it can be extended to the one where the top layer has more than one BS, as discussed in the beginning of section 2.3. In addition, consider that we have the optimal solution for the $(m-1)$-layer IbHWS $\mathcal{A}'$ defined by

$$\mathcal{A}' = \mathcal{A} \setminus a_{m-1} = \{a_0, a_1, a_{m-2}\}. \tag{39}$$

Let E_{N-2} denote the energy consumption of the solution for $\mathcal{A}'$. If we create $\mathcal{A}$ by adding a_{m-1} to $\mathcal{A}'$, we have to decide whether $b_{0,m-1}$ should be *on* or *off*. As in two-layer case, we need to compare the minimum energy required by the network when $b_{0,m-1}$ is *off* against the minimum one required when $b_{0,m-1}$ is *on*. We proceed to find these two values.

If $b_{0,m-1}$ is *off*, the overall network consumption E_{N-3} becomes

$$E_{N-3} = E_{N-2} + \hat{E}_{\text{off}}(b_{0,m-1}). \tag{40}$$

If $b_{0,m-1}$ is *on*, the network energy consumption must increase by at least $\hat{E}_{\text{on,min}}(b_{0,m-1})$. Therefore, in this case,

$$E_{N-3} \geq E_{N-2} + \hat{E}_{\text{on,min}}(b_{0,m-1}). \tag{41}$$

When $E_{\text{min}}(b_{0,m-1})$ is *on* and all locations are served by the BSs in $\mathcal{A}'$, there are four types of actions that could be executed to reduce the energy consumption E_{N-3}.

1. **Action Type 1:** Use the BS in the top layer, i.e., $b_{0,m-1}$, to serve a particular location u instead of $\mathcal{A}'$. This is equivalent to Action Type 1 for the two-layer case.

2. **Action Type 2:** Use $b_{0,m-1}$ to serve all locations u originally served by a particular BS $b_{i,k}$, where $k < m-1$ (i.e., for some $a_k \in \mathcal{A}'$). This is equivalent to Action Type 2 for the two-layer case.

3. **Action Type 3:** For a location u, switch the serving BS to another BS $b_{i,k}$, where $k < m-1$.

4. **Action Type 4:** For all locations u currently served by the same BS, switch the serving BS to another BS $b_{i,k}$, where $k < m-1$.

Remark 10: Applying any Action Type 3 or Action Type 4 to the optimal solution of $\mathcal{A}'$ does not lead to an immediate reduction in the energy consumption E_{N-3}.

Proof: Assume that any Action Type 3 or Action Type 4 can lead to a reduction in the energy consumption E_{N-3}. Thus, any of them could also be applied to the optimal solution of $\mathcal{A}'$ and achieve a reduction in the energy consumption of $\mathcal{A}'$. Since the energy consumption of $\mathcal{A}'$ is already at the minimum, it follows that no action exists that can be applied to further reduce the energy consumption of $\mathcal{A}'$. Therefore, no Action Type 3 or Action Type 4 can lead to an immediate reduction of the energy consumption of $\mathcal{A}'$ or $\mathcal{A}$.

■

As a result of the previous remark, we focus our attention on Action Type 1 and Action Type 2. These two actions are equivalent to the Action Type 1 and Action Type 2 of the two-layer case; hence, we can use the following mapping to avoid repeating all the formulation presented in the two-layer case.

- In the two-layer case, all the active BSs initially serving users belonged to layer 0, i.e., they were of the form $b_{i,0}$. In the m-layer case, they have the form $b_{i,k}$, where $k < m-1$.

- In the two-layer case, the effect of both actions leads to $b_{0,1}$ becoming the serving BS for certain locations u. In the m-layer case, the effect of both actions leads to $b_{0,m-1}$ becoming the serving BS for certain locations u.

With this mapping, we can update all the equations developed for the two-layer case, as well as the proposed algorithm, and apply them to obtain the solution for the m-layer case if we are given the solution to the case of $m-1$ layers. Therefore, given any regular IbHWS $\mathcal{A}$, we can iteratively apply the previous formulation to find the solution for increasing number of layers until we reach m, with an overall complexity $O(mp\log(p))$.

2.4 Performance Evaluation

In this section, we evaluate the performance of the proposed energy-saving algorithm for the regular IbHWS.

2.4.1 Simulation Setup

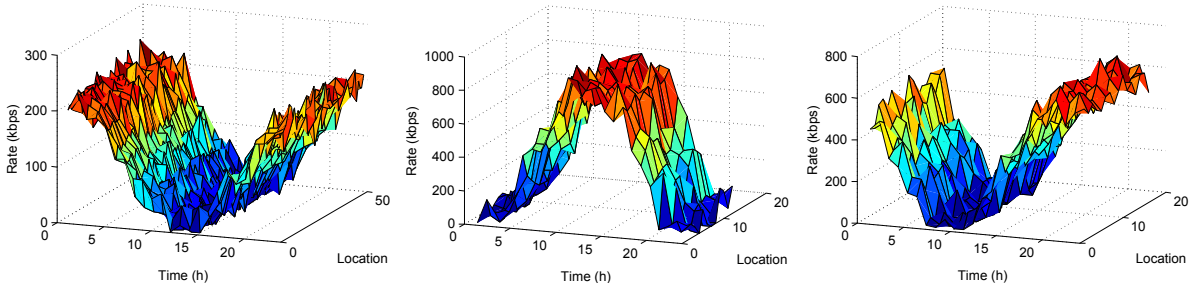
Simulation parameters for the IbHWS are shown in Table 1. BSs per layer and locations are uniformly distributed across the coverage area of interest Ξ.

Table 1: Simulation parameters for multi-layer on-off and cell-association policy.

Parameter	**Value**
Bandwidth	3.84 MHz
BS max. bit rate	14.4 Mbps
QoS rates	[1,20,80,200]*15kbps
Time intervals	24 (1 hr each)
Total coverage radius	2500 m
Number of locations	500
Altitude of locations	1.5 m
Number of layers	3
Type of BSs (per layer)	[macro,pico,pico]
Number of BSs (per layer)	[6,25,100]
Altitude of BSs (per layer)	[25,20,10]m

The IbHWS is set to satisfy the capacity and coverage needs by using three layers. Traffic demands (TDs) are generated so that the second layer (L2) is able to satisfy the peak TD at any time and location, providing the minimum capacity required. To capture the spatio-temporal variability, each L2 BS is set to experience the peak TD at different times, as shown in Figure 7.

The previous formulation also has consequences in the other layers. L1 cannot satisfy all TDs since it has fewer BSs than L2. Consequently, L1 is meant for coverage. On the other hand, L3 has excess capacity to satisfy all TDs. As such, it is meant to enhance capacity. Therefore, this formulation allows to capture operators' objectives of satisfying and enhancing capacity and coverage.



(a) Average load for BS1 in L2. (b) Average load for BS2 in L2. (c) Average load for BS3 in L2.

Figure 7: Average load for base stations in L2.

In Table 2, the parameters of the BS components are listed. We consider that the BSs in L3 require no A/C because of a low expected amount of power consumption. The interconnection among these elements is shown in Figure 8. E_{RF} is the output energy of the feeder. The energy received at the input of the feeder comes from the PA, whose power is mainly drawn from the PS. The energy of the BBuP is also provided by the PS. The A/C needs to compensate the energy dissipated by all the previous elements, and the energy it needs is drawn from the PS. Based on these relationships and the energy models, we obtain the energy used by each BS.

Table 2: Base station components and parameters, per layer.

Component	P_{off}**(W)**	$P_{\text{on,min}}$**(W)**	α	$P_{\text{on,max}}$**(W)**
Feeder	[0,0,0]	[0,0,0]	[1,0,0]*0.63	-
PA	[1,1,1]*0.25	[4,1,0.25]	[2,1,1]*0.2	-
BBuP	[1,1,1]	[4,2,2]	-	[2,1,1]*6
A/C	[1,1,-]*0.9	[1,1,-]*0.9	[3,3,-]	-
PS	[1,1,1]*0.1	[1,1,1]*0.1	[1,1,1]*0.9	-

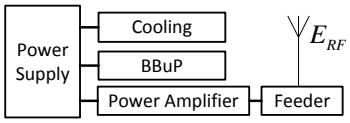


Figure 8: Interconnections of base station components.

For the path loss, we use the following 3GPP models for heterogeneous networks in outdoor scenarios (distance D (in km))[40]:

$$\text{(macro)PL} = 128.1 + 37.6\log(\text{D}), \tag{42}$$

$$\text{(pico)PL} = 140.7 + 36.7\log(\text{D}). \tag{43}$$

2.4.2 Simulation Results

The amount of energy savings that can be achieved compared to an always-on network depends on the specific layout of the network, locations, and traffic demands. First, we present the results obtained from a single scenario generated with the parameters described in Section 2.4.1. Then, we provide values for average energy savings across multiple scenarios.

The results of applying our energy-saving scheme to a scenario generated with the parameters previously described are shown in Figure 9 and Figure 10. In Figure 9, we observe that the energy savings throughout a day were between 34% and 39%. In addition,

the achievable energy savings are inversely related to the overall network load. This behavior can be explained as follows. During a period of low network load, the number of underutilized BSs is generally larger than during a period of high network load. Therefore, for a low network load, there is a higher probability that the BSs can be turned off and, thus, achieve the additional energy savings described by Eq. (33).

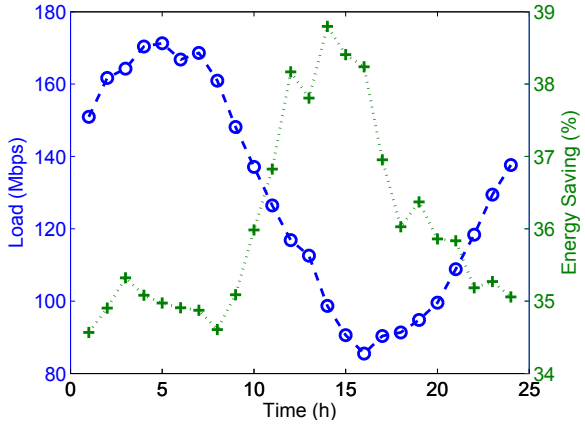
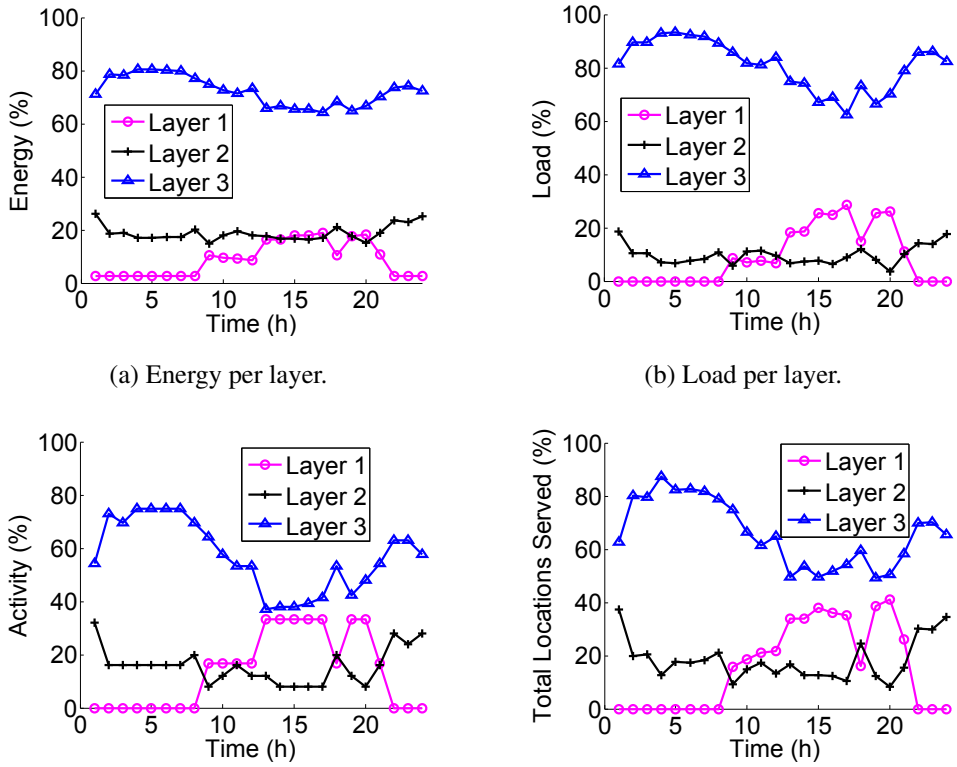


Figure 9: Hourly network load and energy savings.

The effect of our energy-saving algorithm on the load, activity, and the energy consumption across the different layers is depicted in Figure 10. We define the activity of layer i during time interval Δt_j as

$$\text{Activity}(i, \Delta j) = \frac{\sum_{b \in a_i} \zeta(b, \Delta t_j)}{|a_i|}, \tag{44}$$

where $\zeta(b, \Delta t_j)$ is an indicator function equal to 1 when BS b is *on* during time interval Δt_j, and $|a_i|$ denotes the total number of BSs in layer a_i.



(a) Energy per layer.

(b) Load per layer.

(c) Activity per layer.

(d) Locations served per layer.

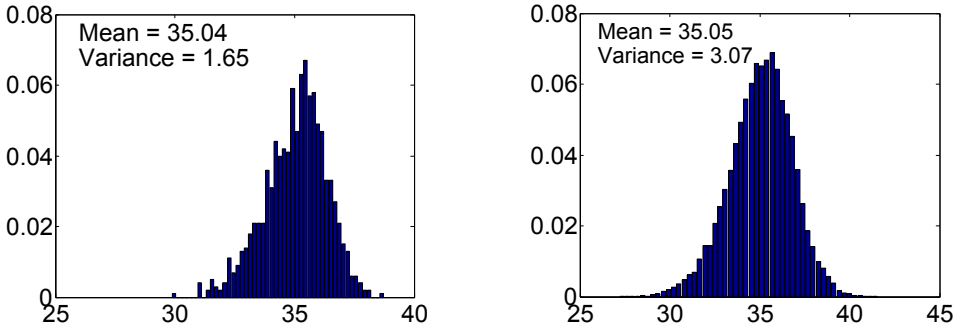
Figure 10: Load, energy, activity, and locations served per layer.

In Figure 10, we observe that L3 handles between 62% to 93% of the traffic throughout the day, while consuming between 64% to 80% of the total network energy. Even though the larger cells (i.e., the ones of BSs in L1 and L2) in general do not handle as much traffic as the ones of L3, the amount of traffic that they manage increases during the time interval 9h-21h, i.e., when the network load is low-to-medium. We also note that the level of activity for L1 and L2 tends to be relatively stable during time intervals of at least 3 hours, and the level of activity of L3 shows more variation, albeit of small amplitude. In addition, we observe that during the time interval 10h-17h, which corresponds in Figure 9 to the highest energy savings, the activity of L3 decreases significantly (i.e., L3 BSs are turned off), and the activity of L1 increases (i.e., L1 BSs are turned on). This behavior suggests that the high energy savings are strongly related to the de-activation of underutilized L3 BSs.

To further explore the effect of our algorithm on the cell-association policy, we obtain

the locations-to-layer association, as shown in Figure 10d. We observe that the increment of the load managed by L1 in the time interval 9h-21h, as shown in Figure 10b, results from L1 becoming the serving layer for an increased number of locations. On the other hand, the relatively stable activity of L2 during the intervals 2h-7h and 13h-17h, as shown in Figure 10c, still corresponds to changes in the cell-association policies during these intervals, as observed in Figure 10d. In general, variations in the cell-association policy during the times of stable activity also occur in the other layers.

To evaluate the overall performance of our energy-saving algorithm, we applied it to 1000 different scenarios generated using the parameters shown in Table 1 and Table 2. Figure 11a shows the probability density function (PDF) for the energy savings. In particular, Figure 11a shows the PDF of the daily energy savings across all scenarios. In this case, the mean is 35.04% and the variance is 1.65. To further analyze how the energy savings behave throughout the day for the different scenarios, we obtain the PDF of the energy savings across all time intervals, as shown in Figure 11b. In this case, the mean is 35.05% and the variance is 3.07. Even though the variance in this case is higher by 1.42 than the one in Figure 11a, the energy savings only differ by 0.01%. This result indicates that the average energy savings across all time intervals do not differ significantly from the average energy savings across multiple scenarios.



(a) Energy savings - all scenarios.

(b) Energy savings - all time intervals.

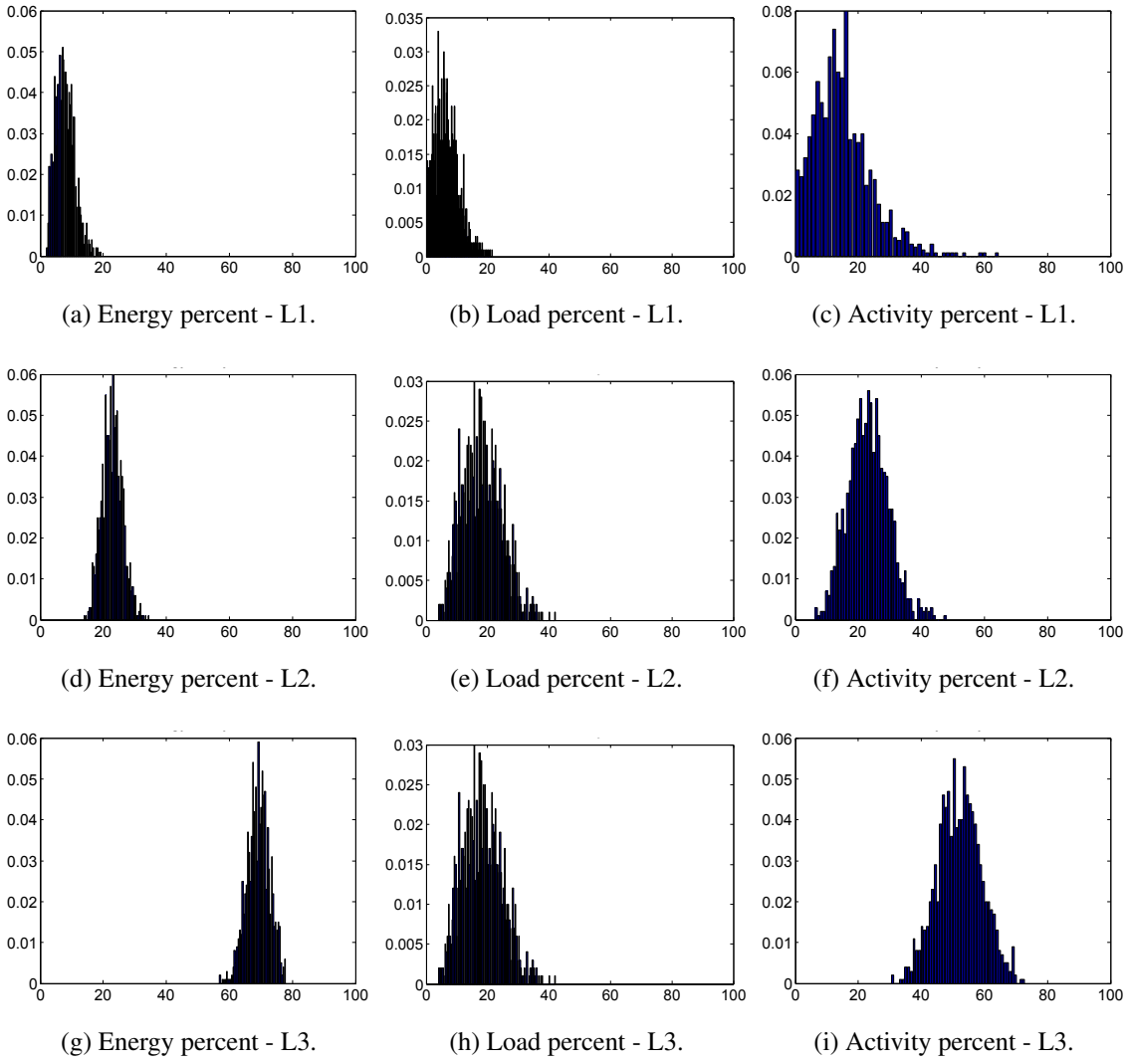
Figure 11: Probability density functions of energy savings.

We also obtain the per-layer PDF for the energy, load, and activity across all scenarios,

as shown in Figure 12. We define the activity of a layer i as

$$\text{Activity}(i) = \frac{\sum_{\Delta t_j} \sum_{b \in a_i} \zeta(b, \Delta t_j)}{\sum_{\Delta t_j} |a_i|}, \tag{45}$$

where $\zeta(b, \Delta t_j)$ is an indicator function equal to 1 when BS b is *on* during the time interval Δt_j, and $|a_i|$ denotes the total number of BSs in layer a_i.



(a) Energy percent - L1. (b) Load percent - L1. (c) Activity percent - L1.

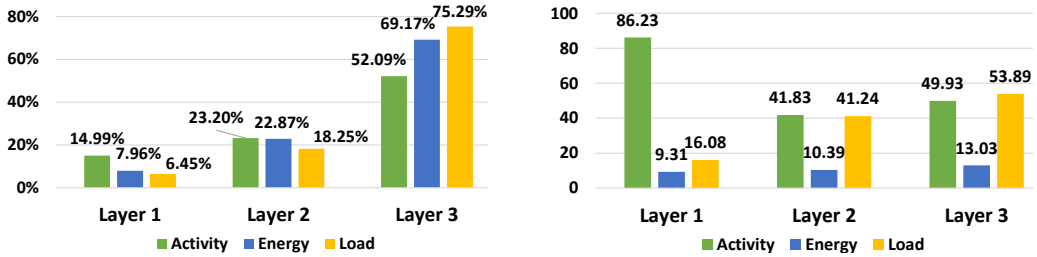
(d) Energy percent - L2. (e) Load percent - L2. (f) Activity percent - L2.

(g) Energy percent - L3. (h) Load percent - L3. (i) Activity percent - L3.

Figure 12: Per-layer probability density functions of energy, load, and activity.

Figure 13 shows the mean and the variance for these PDFs. In terms of the mean values, shown in Figure 13a, we observe that L3 has an activity of around 50%, but manages 75%

of the network load, while consuming no more than 70% of the overall network energy. On the other hand, L2 shows a direct relationship among its activity, energy consumption, and managed load. In contrast to L3, the L1 activity is around 15%; however, L1 only consumes 8% of the energy and handles 6% of the network load. These results highlight the relevance of small cells (i.e., L3) and the non-trivial role of larger cells (i.e., L1 and L2) in achieving energy-efficient networks.



(a) Mean value for activity, energy and load percent.

(b) Variance value for activity, energy and load Percent.

Figure 13: Per-layer statistics.

Figure 13b shows that the variance of the energy consumption is lower than that of the activity and the load and is similar across layers. The high values of the activity and load variance indicate that the per-layer activity and load are highly dependent on the particular IbHWS. On the other hand, the low variance of the energy consumption indicates a low dependency of the energy consumption per layer on the particular IbHWS.

2.5 Conclusions

In this chapter, we have analyzed the energy consumption in HetNets. Specifically, we considered heterogeneous networks composed of multiple layers, as well as the dependence of the energy consumption on the spatio-temporal variations of traffic demands and on the internal hardware components of the BSs. Considering these elements, we studied the energy minimization problem in HetNets, showed a mapping to a 0-1 Knapsack-like problem, and identified the conditions required for a direct mapping to the classical 0-1

Knapsack problem. Given the important differences between the general energy minimization problem and the classical 0-1 Knapsack problem, we developed an efficient algorithm to minimize the energy consumption by adjusting the cell-association and on-off policies of the BSs. We showed that our algorithm is directly applicable to the two-layer case and extendable to the m-layer case. We evaluated the performance of the proposed algorithm and obtained energy savings of 35% across a wide range of scenarios. Throughout the day, these energy savings, in general, vary inversely to the overall network load. Furthermore, we identified that small cells play a key role in not only improving the capacity, but also in increasing the energy efficiency of the network by consuming 69% of the energy while handling 75% of the traffic. However, contrary to the belief that small cells are always the most energy-efficient solution, we found that larger cells have an important role in energy-efficient deployments, particularly during low and medium network load periods. During such periods, the energy consumption of the small cells and their underutilization calls for their de-activation and the utilization of larger cells instead. More importantly, we have shown that the cell-association and on-off policies should be jointly adjusted according to the actual network deployment, the energy efficiency of the BSs, and the traffic dynamics to achieve a highly energy-efficient network operation.

CHAPTER 3

ENERGY-EFFICIENT MULTI-STREAM CARRIER AGGREGATION IN HETNETS

Multi-stream carrier aggregation (MSCA) has been recently proposed as a mechanism to increase the amount of bandwidth available to users in HetNets. Nevertheless, existing work has focused only on maximizing the network capacity and fairness, without considering the energy efficiency of MSCA. In this chapter, the use of MSCA to minimize the energy consumption in a multi-layer HetNet is studied. The convexity of the energy minimization problem is examined, leading to the need of a quasiconvex relaxation. With this approximation, a simple algorithm is designed to solve the energy minimization problem and obtain an optimum cell-association policy. Since the operators are generally interested in a balance between the energy minimization and capacity maximization, such multi-objective optimization is studied here. We show that the two aforementioned conflicting objectives can be jointly analyzed and solved through scalarization, even though the energy minimization has a quasiconvex objective function, and not a convex one. Performance evaluation is provided to identify the achievable energy savings of our proposed algorithm and to characterize the trade-offs between the energy minimization and capacity optimization in a multi-layer HetNet that supports MSCA.

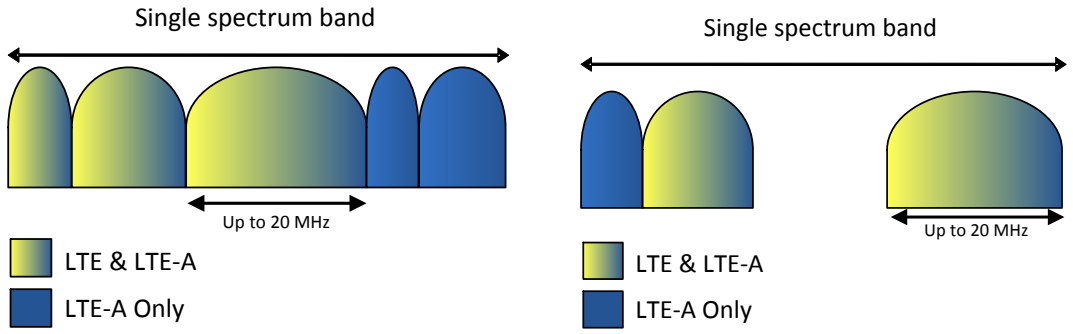
3.1 Motivation and Related Work

One of the most effective methods to improve the network performance is to increase the amount of utilized bandwidth. Therefore, to meet the requirements of IMT-Advanced, as well as those of the operators, LTE-Advanced considers the use of bandwidths of up to 100MHz in several frequency bands. These bands are set by the ITU for IMT, and include the following [41]: 450-470MHz, 698-960MHz, 1710-2025MHz, 2110-2200MHz, 2300-2400MHz, 2500-2690MHz, 3400-3600MHz. LTE-Advanced tries to exploit as much as

possible the flexibility of supporting multiple frequency bands through the use of carrier aggregation [42][43].

Carrier aggregation (CA) consists of grouping several component carriers (CC) to achieve wider transmission bandwidths. An LTE-Advanced device can aggregate up to five CCs, each of up to 20 MHz. With the largest configuration, this implies a total bandwidth of 100MHz. To support backward compatibility with LTE devices, each of the CCs shall be configured as a typical LTE carrier. Therefore, any of the CCs used for CA should also be accessible to LTE UEs. Nevertheless, mechanisms, such as barring [44], already exist to prevent LTE UEs from camping on specific CCs. This way, operators have the flexibility of adjusting the characteristics of the CCs to support a mixture of LTE and LTE-Advanced devices.

LTE-Advanced supports three schemes of CA [45][46]. The first two can be grouped as intra-band carrier aggregation, as shown in Figure 14. The basic scheme of CA occurs when contiguous CCs are aggregated, as shown in Figure 14a. It depicts the aggregation of five CCs of different bandwidths, where the two rightmost are used exclusively by LTE-Advanced devices and the rest is shared among LTE and LTE-Advanced devices. While this scenario is the easiest to implement from the technical point of view, operators do not always have enough contiguous spectrum to perform this type of deployment.



(a) Contiguous.

(b) Non-contiguous.

Figure 14: Intra-band carrier aggregation.

Therefore, non-contiguous CA is also supported. In this case, the CCs may belong to the same or different spectrum bands, also called intra-band non-contiguous CA and inter-band CA, respectively. These two scenarios are depicted in Figure 14b and Figure 15. These two types of CA are extremely useful to the operators who have fragmented spectrum along multiple frequency bands since they allow the operators to effectively reuse those spectrum fragments to provide improved service to their users.

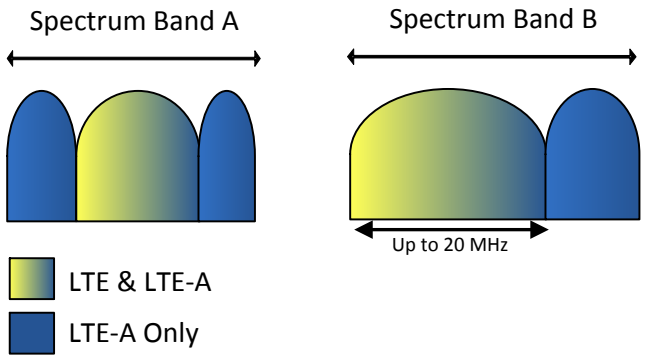


Figure 15: Inter-band carrier aggregation.

To obtain the most benefit from CA, each BS should support the maximum number of CCs[10]. However, in most HetNet scenarios, not all BSs support such configuration. This is mainly due to the hardware limitations, which require costly upgrades.

Multi-stream CA (MSCA), or multi-flow CA, has been recently proposed as an alternative method to address this problem [47][48][49]. In MSCA, a UE is able to aggregate CCs belonging to multiple BSs, allowing it to achieve the maximum CA configuration, even if no BS can provide such configuration by itself. In non-MSCA networks, the existing literature has looked at multiple ways of reducing the network energy consumption. In [50], the use of lean carriers with reduced signaling overhead is proposed. By reducing the signaling overhead, the BS can go into micro-sleep more frequently. The concept of adjusting the cell-association policies and, therefore, the load across BSs has also been

[10]The current version of the standard specifies the maximum configuration to be five CCs.

proposed separately for energy minimization [35] and user fairness [38] [51] [52]. Furthermore, cooperation among BSs has been utilized to minimize the energy consumption by coordinating the scheduling and power control mechanisms [53] [54] and the on-off policies [55] [56]. Compared to the literature on non-MSCA, existing work on MSCA-enabled networks has focused on maximizing the network capacity [57][58], but almost none has been done on analyzing energy-efficient methods of exploiting MSCA [59].

The focus of our work is on designing new methods of exploiting MSCA to improve the energy efficiency in multi-layer HetNets. In particular, we show that the energy minimization problem in MSCA-enabled networks is a non-convex optimization. Nevertheless, such problem can be approximated through a generalized linear-fractional program. Using this approximation, we develop a simple algorithm to solve such problem by applying a bisection method that solves a convex feasibility problem at each step, until a precision tolerance is met. Since the operators are typically interested not only in minimizing the energy consumption, but also in maximizing the network capacity, we analyze these problems jointly as a multi-objective optimization. Based on the analysis done for the energy minimization problem, we provide a solution for the multi-objective one, according to the priority assigned by the operators to each objective. Moreover, we show that an explicit analytical expression for the UE-to-CC association policy can be obtained without the need of solving the multi-objective optimization problem.

The rest of this chapter is organized as follows. We present the network architecture and BS energy model in Sections 3.2.1 and 3.2.2, respectively. In Section 3.3, we develop the energy- and capacity-aware mechanisms of load balancing that exploit the use of MSCA. In particular, in Section 3.3.1, we focus on the single objective of minimizing the network energy consumption. Then, the energy minimization and capacity maximization are jointly analyzed in Section 3.3.2. Simulation results showing the performance of our load-balancing mechanisms are presented in Section 3.4. Finally, the conclusions are presented in Section 3.5.

3.2 System Model

3.2.1 Network Architecture

As discussed in Section 2.2.1, we classify a IbHWS as either a general or a regular one. In this chapter, we focus on a general IbHWS where the UE and the BSs support MSCA. An example of such network is depicted in Figure 16, where the UE applies MSCA by aggregating three CCs belonging to different BSs in a three-layer general IbHWS.

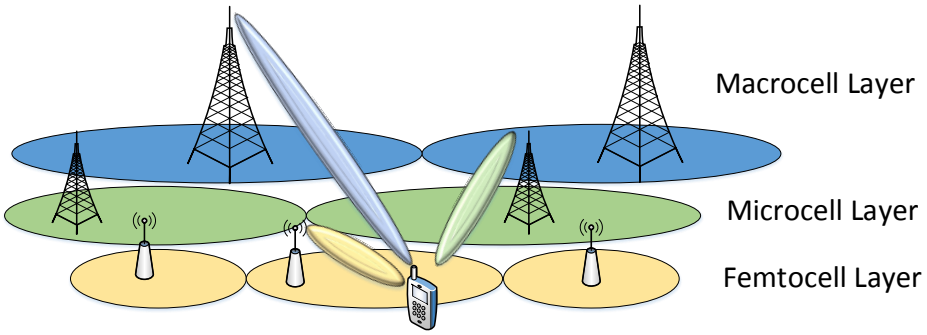


Figure 16: MSCA in a general IbHWS.

From Figure 16, we observe that the UE applies MSCA by connecting to CCs that belong to BSs of different layers. This behavior follows from the fact that the BSs in the same layer are typically assigned the same frequency bands, i.e., they are configured with a frequency reuse factor of one. Therefore, if a UE were to connect to multiple BSs of the same layer, it would have to utilize advanced intercell interference cancellation (ICIC) techniques to recover the signal from each BS [60][61][62]. While not impossible, such functionality is not within the scope of MSCA; rather, it is included in cooperative multipoint transmission and reception (CoMP) [63][64]. CoMP requires not only advanced ICIC at the UE, but also a significant amount of coordination between the BSs, so that their transmission scheme is suitable for the application of advanced ICIC at the UE [65][66][67]. Compared to CoMP, the overhead associated with MSCA is significantly smaller.

Similar to the approach followed in Section 2.2.1, we consider that the network serves a set of users $\mathcal{U}$ and that each user can generate sessions that have different QoS requirements $q \in \mathcal{Q}$. These are defined in terms of bit rates, to capture the non-linear relationship between

bit rate and power. In addition, we consider that the each UE is capable of partitioning every session across all of its active CCs. Such assumption is justified by the fact that, if we were to consider that a session can only be served through a single CC, the overall problem would be reduced to a flow-to-CC association problem, which is equivalent to the traditional UE-to-CC cell-association problem when no MSCA is considered. Finally, we also consider that the CCs have full buffers.

3.2.2 Base Station Energy Model

In terms of the energy consumption of the BS, we consider that each BS b has a set $\mathcal{S}_b$ of CCs and that the energy consumption of each CC follows the model developed and analyzed in Section 2.2.3. Based on such model, the total dynamic energy consumption of a CC $s \in \mathcal{S}_b$ that is *on* depends on the total bit rate served by the CC, the RF output power required to satisfy the QoS requirements requested by the UEs connected to the CC, and the internal components associated with the operation of the CC and their interconnection. In particular, such energy can be described as a linear function of the total RF power that the CC needs to radiate to satisfy its UE QoS requirements.

3.3 Energy- and Capacity-Aware Load Balancing

In a traditional macrocell-only cellular network, the default user cell-association policy is to assign the user to the BS that provides the maximum SINR, i.e., a max-SINR policy. Since the energy consumption model of all the BSs in such network is approximately the same, following a max-SINR policy causes the UE to attach to the BS that not only maximizes the coverage probability, but also minimizes the amount of energy consumed by the BS to serve such user and by the user to communicate with the BS. However, with the introduction of HetNets, the max-SINR policy is no longer the optimum policy to achieve such objectives.

In a HetNet, the macrocells transmit at a much higher power than the small cells. Therefore, the max-SINR policy tends to favor the association towards the macrocells over small cells. As a result, a UE close to a small cell may still connect to a macrocell even if

- communicating with the macrocell requires more energy, in uplink or downlink, than communicating with the small cell or
- the macrocell is overloaded, and the small cell resources are being underutilized. Such situation may cause the user QoS requirements to not be satisfied by the macrocell, even though the small cell could have done so.

Moreover, even if a user is capable of following a cell-association policy different from the max-SINR one, the fact that it can only connect to a single BS means that

- no single cell may have enough capacity to satisfy the downlink and uplink QoS requirements or
- a user may connect to a cell that can satisfy the downlink or uplink QoS requirements, but not both or
- the non-linear relationship between the minimum power received and bit rate would require a disproportional amount of power to satisfy the QoS requirements.

Thus, MSCA can address the aforementioned issues, given that we design mechanisms to balance the load across small cells and macrocells while accounting for the energy consumption and the capacity of each one. The design of such mechanisms for the downlink is the focus of this chapter. The uplink could be analyzed by following a similar approach, assuming that an accurate model for the UE total dynamic energy consumption is available.

From now on, we will utilize the notation $\mathrm{CC}_{j,k}$ to denote CC k of BS j. For UE i and $\mathrm{CC}_{j,k}$, the maximum spectral efficiency $\theta_{i,j,k}$ of the communication link is a logarithmic function of the SINR:

$$\theta_{i,j,k} = \beta \log_2 \left(1 + \underbrace{h_{i,j,k} \frac{P_{i,j,k}}{\eta_{i,j,k}}}_{\text{SINR}} \right), \tag{46}$$

where $P_{i,j,k}$ is the RF output power used by $\mathrm{CC}_{j,k}$ on the resources assigned to UE i, $h_{i,j,k}$ is the channel gain between UE i and $\mathrm{CC}_{j,k}$, $\eta_{i,j,k}$ represents the noise plus interference

experienced by UE i when connected to $\mathrm{CC}_{j,k}$, and $0 < \beta < 1$ is a factor that accounts for the proximity to the Shannon channel capacity that the modulation and coding scheme (MCS) can achieve[11]. The factor $h_{i,j,k}$ includes the path loss, fading, and shadowing effects. By considering a large time scale for the association between a user and a CC, the short-term channel dynamics, such as fast fading, can be averaged out, allowing us to consider the SINR and the spectral efficiency as constants during the association duration.

Even though the maximum spectral efficiency $\theta_{i,j,k}$ is a good metric for the quality of the channel between user i and $\mathrm{CC}_{j,k}$, the overall bit rate is the metric of interest when determining if the QoS is satisfied. The bit rate achieved over the aforementioned channel depends not only on $\theta_{i,j,k}$, but also on the amount of resources assigned to such channel by the BS. Particularly, for user i, $\mathrm{CC}_{j,k}$ with bandwidth $\rho_{j,k}$, the maximum bit rate $\tilde{\theta}_{i,j,k}$ over a channel that has a maximum spectral efficiency $\theta_{i,j,k}$ is

$$\begin{aligned}\tilde{\theta}_{i,j,k} &= \rho_{j,k} y_{i,j,k} \theta_{i,j,k} \\ &= \rho_{j,k} y_{i,j,k} \beta \log_2\left(1 + h_{i,j,k}\frac{P_{i,j,k}}{\eta_{i,j,k}}\right),\end{aligned} \tag{47}$$

where the factor $0 \leq y_{i,j,k} \leq 1$ represents the fraction of resources reserved for user i by $\mathrm{CC}_{j,k}$, and $\theta_{i,j,k}$ is obtained from Eq. (46). By considering that the resource allocation is performed within the coherence time of the channel, the latter can be considered static during every allocation period. Such assumption is valid for low-mobility scenarios.

Typically, $y_{i,j,k} < 1$ since the BS serves more than one user; therefore, it must allocate the limited resources among those users. As a result, the achievable rate of a user depends not only on the channel quality towards a particular CC, but also on the number of other users associated with such CC and the resource allocation policy followed. The latter depends directly on how much bit rate, i.e., load, each UE requests from each CC.

In Section 3.3.1, we focus on finding an optimal load-balancing and cell-association

[11] In LTE, β is approximately 0.75 [68]

policy that minimizes the energy consumption of the network. Then, in Section 3.3.2, we utilize the results form Section 3.3.1 to develop an optimal load-balancing and cell-association policy capable of addressing the conflicting objectives of minimizing the network energy consumption and maximizing its capacity.

3.3.1 Load Balancing for Energy Minimization

Based on the QoS requirements of all the sessions that a user i needs to support, a total bit rate r_i can be computed for such user. By using MSCA, r_i can be split across all the CCs to which the UE is capable of connecting. If UE i requests a fraction $0 \leq \xi_{i,j,k} \leq 1$ of r_i to $\mathrm{CC}_{j,k}$, then the following relationship must hold:

$$\begin{aligned} \xi_{i,j,k} r_i &\leq \tilde{\theta}_{i,j,k} \\ &= \rho_{j,k} y_{i,j,k} \beta \log_2 \left(1 + h_{i,j,k} \frac{P_{i,j,k}}{\eta_{i,j,k}}\right), \end{aligned} \tag{48}$$

where $\tilde{\theta}_{i,j,k}$ is obtained from Eq. (47). The above expression conveys that the amount of bit rate requested by any UE to any CC should not exceed the capacity of the channel between them for a given amount of bandwidth and power allocated to the user. Typically, the above inequality can be treated as an equality, since there is no benefit to the user or BS to have an underutilized channel. Assuming that $y_{i,j,k} > 0$, i.e., that $\mathrm{CC}_{j,k}$ is assigning a non-zero amount of bandwidth to user i, the amount of output RF power at the antenna of $\mathrm{CC}_{j,k}$ for user i can be obtained as follows,

$$\begin{aligned} P_{i,j,k} &= \left[\exp\left(\frac{1}{\beta} \frac{\xi_{i,j,k} r_i}{\rho_{j,k} y_{i,j,k}} \ln(2)\right) - 1\right] \frac{\eta_{i,j,k}}{h_{i,j,k}} \\ &= \frac{\eta_{i,j,k}}{h_{i,j,k}} \exp\left(\frac{1}{\beta} \frac{\xi_{i,j,k} r_i}{\rho_{j,k} y_{i,j,k}} \ln(2)\right) - \frac{\eta_{i,j,k}}{h_{i,j,k}}. \end{aligned} \tag{49}$$

As discussed in Section 3.2.2, the total dynamic power $\hat{P}_{\mathrm{on,dyn}}$ consumed by $\mathrm{CC}_{j,k}$ to output the RF power required to satisfy the QoS requirements of the UEs is a linear function of

$P_{i,j,k}, \forall i, j, k$:

$$\begin{aligned}\hat{P}_{\text{on,dyn}}(\text{CC}_{j,k}) &= w_{j,k} \sum_i P_{i,j,k} \\ &= w_{j,k} \sum_i \frac{\eta_{i,j,k}}{h_{i,j,k}} \exp\left(\frac{1}{\beta} \frac{\xi_{i,j,k} r_i}{\rho_{j,k} y_{i,j,k}} \ln(2)\right) - w_{j,k} \sum_i \frac{\eta_{i,j,k}}{h_{i,j,k}},\end{aligned} \tag{50}$$

where $w_{j,k}$, a unique weight for every $\text{CC}_{j,k}$, depends on the actual components associated with the operation of $\text{CC}_{j,k}$ and their interconnection. Based on the expression above, the network energy minimization problem can be described as

$$\text{minimize} \quad \sum_j \sum_k \hat{P}_{\text{on,dyn}}(\text{CC}_{j,k}), \tag{51a}$$

$$\text{subject to} \quad \sum_j \sum_k \xi_{i,j,k} = 1, \qquad \forall i, \tag{51b}$$

$$\sum_i y_{i,j,k} \leq 1, \qquad \forall j, k, \tag{51c}$$

$$\xi_{i,j,k} \geq 0, \qquad \forall i, j, k, \tag{51d}$$

$$y_{i,j,k} \geq 0, \qquad \forall i, j, k, \tag{51e}$$

$$r_i \xi_{i,j,k} - \rho_{j,k} \theta_{\max} y_{i,j,k} \leq 0, \qquad \forall i, j, k, \tag{51f}$$

where $\theta_{\max}$ denotes the maximum spectral efficiency supported by the network[12]. Constraint (51b) indicates that the UE total QoS requirement r_i must be satisfied with equality. Constraint (51c) indicates that $\text{CC}_{j,k}$ cannot allocate more bandwidth than it has available. Constraints (51d) and (51e) indicate that the allocation variables $\xi_{i,j,k}$ and $y_{i,j,k}$ are non-negative. Constraint (51f) indicates that every channel should operate within the maximum spectral efficiency supported by the network. All the constraints in the optimization problem (51) are linear expressions.

[12] In LTE, $\theta_{\max}$ is approximately 4.8 bits/sec/Hz [68].

The optimization problem (51) is equivalent to

$$\text{minimize} \qquad \sum_i \sum_j \sum_k w_{j,k} \frac{\eta_{i,j,k}}{h_{i,j,k}} \exp\left(\frac{1}{\beta} \frac{\xi_{i,j,k} r_i}{\rho_{j,k} y_{i,j,k}} \ln(2)\right), \tag{52a}$$

$$\text{subject to} \qquad \text{Constraints (51b)-(51f)}, \tag{52b}$$

where the optimization variables are all the $\xi_{i,j,k}$ and $y_{i,j,k}$, and we drop the last term of Eq. (50) because it is a constant that does not affect the solution of the problem. It is important to highlight that objective function (52a) does not allow for $y_{i,j,k} = 0$, even though constraint (51e) does. We will later see that the objective function can be further approximated to allow $y_{i,j,k} = 0$.

Since the factors $w_{j,k}$, $\eta_{i,j,k}$, and $h_{i,j,k}$ are positive constants, the optimization problem (52) can be further rewritten as

$$\text{minimize} \qquad \sum_i \sum_j \sum_k \exp\left(\ln\left(w_{j,k} \frac{\eta_{i,j,k}}{h_{i,j,k}}\right) + \frac{1}{\beta} \frac{\xi_{i,j,k} r_i}{\rho_{j,k} y_{i,j,k}} \ln(2)\right), \tag{53a}$$

$$\text{subject to} \qquad \text{Constraints (51b)-(51f)}. \tag{53b}$$

We now analyze the convexity of the optimization problem (53). Consider a single term of the summation in the objective function. Any such term is an exponential function whose argument is a linear-fractional function. Since a linear-fractional function is quasiconvex, and the exponential function is a non-decreasing function, it follows that each term in the above summation is also quasiconvex. Nevertheless, while convexity is preserved by a nonnegative weighted sum operation, quasiconvexity may not be [69]. Therefore, the optimization problem (53) needs to be reformulated before any convex or quasiconvex optimization technique can be applied.

First, since the logarithmic function is a monotonically increasing function, we apply it

to the objective function (53a) to create an equivalent optimization problem:

$$\text{minimize} \qquad \ln\left(\sum_i \sum_j \sum_k \exp\left(\ln\left(w_{j,k}\frac{\eta_{i,j,k}}{h_{i,j,k}}\right) + \frac{1}{\beta}\frac{\xi_{i,j,k}r_i}{\rho_{j,k}y_{i,j,k}}\ln(2)\right)\right), \tag{54a}$$

$$\text{subject to} \qquad \text{Constraints (51b)-(51f).} \tag{54b}$$

Second, we exploit the log-sum-exp approximation

$$\max\{x_1,\ldots,x_n\} \leq \ln\left(e^{x_1}+\cdots+e^{x_n}\right) \leq \max\{x_1,\ldots,x_n\} + \ln n \tag{55}$$

to apply it to the optimization problem (54), which then becomes

$$\text{minimize} \qquad \max_{i,j,k}\left[\ln\left(w_{j,k}\frac{\eta_{i,j,k}}{h_{i,j,k}}\right) + \frac{1}{\beta}\frac{\xi_{i,j,k}r_i}{\rho_{j,k}y_{i,j,k}}\ln(2)\right], \tag{56a}$$

$$\text{subject to} \qquad \text{Constraints (51b)-(51f).} \tag{56b}$$

The objective function (56a) is a maximization of linear-fractional functions and, therefore, of quasiconvex functions. Such formulation is also known in the literature as a generalized linear-fractional program. Since a nonnegative weighted maximum of quasiconvex functions is also quasiconvex, so is the above objective function. We can now apply any general approach for quasiconvex programming. One such approach consists in representing the sublevel sets of the quasiconvex function via a family of convex inequalities, as we will now describe [69].

First, we define $g_0(\xi, y)$ as our current objective function:

$$g_0(\xi, y) = \max_{i,j,k}\left[\ln\left(w_{j,k}\frac{\eta_{i,j,k}}{h_{i,j,k}}\right) + \frac{1}{\beta}\frac{\xi_{i,j,k}r_i}{\rho_{j,k}y_{i,j,k}}\ln(2)\right]. \tag{57}$$

Second, for a given parameter μ, $g_0(\xi, y) \leq \mu$ if and only if

$$\max_{i,j,k} \left[\xi_{i,j,k} r_i \ln(2) - \beta \rho_{j,k} y_{i,j,k} \left[\mu - \ln\left(w_{j,k} \frac{\eta_{i,j,k}}{h_{i,j,k}} \right) \right] \right] \leq 0. \tag{58}$$

If we define the convex function $\vartheta_\mu(\xi, y)$ as

$$\vartheta_\mu(\xi, y) = \max_{i,j,k} \left[\xi_{i,j,k} r_i \ln(2) - \beta \rho_{j,k} y_{i,j,k} \left[\mu - \ln\left(w_{j,k} \frac{\eta_{i,j,k}}{h_{i,j,k}} \right) \right] \right], \tag{59}$$

then

$$g_0(\xi, y) \leq \mu \Leftrightarrow \vartheta_\mu(\xi, y) \leq 0. \tag{60}$$

Therefore, the μ-sublevel set of the quasiconvex function g_0 is the 0-sublevel set of the convex function ϑ_μ.

Let us denote the optimal value of the quasiconvex optimization problem (56) as χ^*. If the feasibility problem

$$\begin{aligned} \text{find} \quad & \xi, y, && \text{(61a)} \\ \text{subject to} \quad & \vartheta_\mu(\xi, y) \leq 0, && \text{(61b)} \\ & \text{Constraints (51b)-(51f)}, && \text{(61c)} \end{aligned}$$

is feasible, then $\chi^* \leq \mu$ and any feasible point (ξ, y) is also a feasible point for the quasiconvex problem (56). If the problem (61) is not feasible, then $\chi^* > \mu$. Problem (61) is a convex feasibility problem. Therefore, we can verify whether χ^* is greater or less than a particular value μ by solving problem (61). Based on this last observation, a simple procedure to find χ^* is designed through a bisection method that solves a convex feasibility problem at every step, as described in Algorithm 4.

In Algorithm 4, assuming that the quasiconvex problem (56) is feasible and that we know an interval $[l_1, l_2]$ that contains the optimal value χ^*, we solve the feasibility problem

Algorithm 4 Bisection method for energy minimization.

given $l_1 \leq \chi^*$, $l_2 \geq \chi^*$, $\epsilon > 0$
repeat
    $\mu = (l_2 + l_1)/2$
    Solve the convex feasibility problem (61)
    if (61) is feasible **then**
        $l_1 = \mu$
    else
        $l_2 = \mu$
    end if
until $l_2 - l_1 \leq \epsilon$

at the midpoint $\mu = (l_1 + l_2)/2$ of such interval by applying any convex optimization technique, e.g., interior-point method. The result of the feasibility problem indicates whether χ^* is in the lower or upper half of the interval, which we then use to update the interval accordingly. The new interval is half the size of the initial one, i.e., it is bisected. This procedure is repeated until the size of the interval satisfies some lower bound ϵ. After m iterations, the size of the interval is $2^{-m}(l_2 - l_1)$. Therefore, the number of iterations required before the algorithm terminates is $\left\lceil \log_2 \left((l_2 - l_1)/\epsilon \right) \right\rceil$.

To apply Algorithm 4, we need the initial interval $[l_1, l_2]$ that is guaranteed to contain the optimal value χ^*. Such interval can be obtained from constraint (61b), as shown below. For such constraint to be satisfied, the following expression must be true:

$$\max_{i,j,k} \left[\xi_{i,j,k} r_i \ln(2) - \beta \rho_{j,k} y_{i,j,k} \left[\mu - \ln\left(w_{j,k} \frac{\eta_{i,j,k}}{h_{i,j,k}} \right) \right] \right] \leq 0, \tag{62}$$

which is equivalent to

$$\xi_{i,j,k} r_i \ln(2) - \beta \rho_{j,k} y_{i,j,k} \left[\mu - \ln\left(w_{j,k} \frac{\eta_{i,j,k}}{h_{i,j,k}} \right) \right] \leq 0, \qquad \forall i, j, k. \tag{63}$$

Since r_i is positive and $\xi_{i,j,k}$ is non-negative, then the first term is also non-negative. Therefore, we need the second term to be non-positive. As a result, there are two possible necessary conditions for the above expression to be satisfied for any given i,j,k:

$$\mu - \ln\left(w_{j,k}\frac{\eta_{i,j,k}}{h_{i,j,k}}\right) > 0 \tag{64}$$

or

$$\xi_{i,j,k} = y_{i,j,k} = 0. \tag{65}$$

From the former, we can now obtain an interval $[l_1, l_2]$ that is guaranteed to include the optimal value χ^*:

$$l_1 = \min_{i,j,k} \ln\left(w_{j,k}\frac{\eta_{i,j,k}}{h_{i,j,k}}\right), \tag{66}$$

$$l_2 = \max_{i,j,k} \ln\left(w_{j,k}\frac{\eta_{i,j,k}}{h_{i,j,k}}\right). \tag{67}$$

With such interval, we can now apply Algorithm 4 to find the optimal value χ^* for the energy minimization problem. Once such value is found, the energy minimization problem can be expressed as a single convex feasibility problem:

$$\begin{aligned}
\text{find} \quad & \xi, y, && && \text{(68a)}\\
\text{subject to} \quad & \xi_{i,j,k} r_i \ln(2) - \beta\rho_{j,k} y_{i,j,k}\left[\chi^* - \ln\left(w_{j,k}\frac{\eta_{i,j,k}}{h_{i,j,k}}\right)\right] \leq 0, && \forall i, j, k, && \text{(68b)}\\
& \text{Constraints (51b)-(51f).} && && \text{(68c)}
\end{aligned}$$

However, the optimum UE-to-CC association policy can be directly obtained from knowing χ^* without the need to solve this last optimization problem. From Eq. (64) and

Eq. (65), we have that

if

$$\chi^* \leq \ln\left(w_{j,k}\frac{\eta_{i,j,k}}{h_{i,j,k}}\right), \tag{69}$$

then

$$\xi_{i,j,k} = y_{i,j,k} = 0. \tag{70}$$

We can state in an equivalent way that the optimum UE-to-CC association policy for energy minimization is for UE i to associate with $\mathrm{CC}_{j,k}$ if and only if

$$\frac{h_{i,j,k}}{\eta_{i,j,k}} > w_{j,k}e^{-\chi^*}. \tag{71}$$

3.3.2 Load Balancing for Joint Energy Minimization and Capacity Maximization

In Section 3.3.1, we analyzed the energy minimization problem in a general HetNet where MSCA is supported. Even though the operators are able to reduce their economic and environmental impact by minimizing the energy consumption, they are typically interested in finding a balance between reducing the energy consumption and maximizing the network capacity. In this section, we analyze how these two conflicting objectives can be addressed jointly.

In a capacity maximization problem, the objective function generally follows the form of

$$f_1(\xi) = \sum_i r_i U\left(\sum_j \sum_k \xi_{i,j,k}\right), \tag{72}$$

or

$$f_2(\xi) = \sum_j U\left(\sum_k \sum_i r_i \xi_{i,j,k}\right), \tag{73}$$

where U is a concave function. A typical approach used in the literature is to consider U to be a logarithmic function. In such case, $f_1(\xi)$ represents a metric of fairness across multiple UEs, i.e., it is better to increase the bit rate of a user that is experiencing a low bit rate than to increase that of a user with an already high bit rate. Similarly, $f_2(\xi)$ represents a metric of load fairness across BSs, i.e., it is better to increase the total load (bit rate) carried by an underloaded BS than to increase the load of a BS that is already carrying a high load. Rather than focusing on a specific case, we will utilize a generic concave function $f_3(\xi)$. For such function, the capacity maximization problem can be expressed as

$$\begin{aligned}
\text{maximize} \quad & f_3(\xi), && && \text{(74a)}\\
\text{subject to} \quad & \sum_j \sum_k \xi_{i,j,k} \geq 1, && \forall i, && \text{(74b)}\\
& \sum_i y_{i,j,k} \leq 1, && \forall j,k, && \text{(74c)}\\
& \xi_{i,j,k} \geq 0, && \forall i,j,k, && \text{(74d)}\\
& y_{i,j,k} \geq 0, && \forall i,j,k, && \text{(74e)}\\
& r_i \xi_{i,j,k} - \rho_{j,k}\theta_{\max} y_{i,j,k} \leq 0, && \forall i,j,k. && \text{(74f)}
\end{aligned}$$

The only difference between the above constraints and the ones of the energy minimization problem is that here, the UE total QoS requirement r_i no longer needs to be satisfied with equality; rather, it is the lower bound, specified by constraint (74b). Therefore, the domain of the energy minimization problem is a subset of the one of the capacity maximization problem. Moreover, any feasible point for the energy minimization problem (51) is also feasible for the capacity maximization problem (74).

We can reformulate the capacity maximization problem as a convex minimization problem:

$$\text{minimize} \qquad f_4(\xi) \equiv -f_3(\xi), \tag{75a}$$

$$\text{subject to} \qquad \text{Constraints (74b)-(74f)}, \tag{75b}$$

where $f_4(\xi)$ represents the new objective function. Since f_4 is the negative of a concave function, it is convex. If we denote by $f_0(\xi, y)$ the objective function of the energy minimization problem, then the problem of jointly minimizing the energy consumption and maximizing the network capacity can be expressed as

$$\text{minimize} \qquad \begin{bmatrix} f_0(\xi, y) \\ f_4(\xi) \end{bmatrix}, \tag{76a}$$

$$\text{subject to} \qquad \text{Constraints (74b)-(74f)}, \tag{76b}$$

i.e., as a *multi-criterion* or *multi-objective* optimization problem. It is important to note that f_0 and f_4 are competing functions, i.e., one of them is minimized at the expense of increasing the other. Because of this competing nature, no single point is capable of jointly achieving the minimum value that f_0 and f_4 could achieve separately. However, since a multi-objective optimization is a vector optimization defined over a cone $K = \mathbb{R}_+^m$ for some $m > 0$, we can scalarize the problem to find Pareto-optimal points for the original problem. Applying scalarization to the optimization problem (76), we obtain

$$\text{minimize} \qquad \upsilon f_0(\xi, y) + (1 - \upsilon) f_4(\xi), \tag{77a}$$

$$\text{subject to} \qquad \text{Constraints (74b)-(74f)}, \tag{77b}$$

where $0 \leq \upsilon \leq 1$ is a parameter that is adjusted to find all the Pareto-optimal points.

Intuitively, υ is selected to indicate the operator's balance point between the energy minimization and the capacity maximization. For υ close to 1, a greater weight is given to the energy minimization. Conversely, for υ close to 0, a greater weight is given to the capacity maximization.

In general, for a given υ, if f_0 and f_4 are convex functions, then the scalarized optimization problem is a convex one. We have shown that f_4 is a convex function, and, in Section 3.3.1, we found that the energy minimization problem can be expressed as the convex feasibility problem (74). The problem with using the objective function of the latter is that, by definition, the objective function of a feasibility problem is a constant independent of the optimization variables. If we were to consider f_0 a constant, then it would have no effect on the solution of the optimization problem (77), i.e., such optimization problem would be reduced to the capacity maximization problem. Therefore, f_0 cannot be directly taken from the convex feasibility problem (74). However, we can obtain an appropriate f_0 from the original formulation of energy minimization problem, as we will now describe.

If we apply the weight factor υ to the objective function of the original formulation of the energy minimization problem in (51), such problem becomes

$$\text{minimize} \qquad \upsilon \sum_{j} \sum_{k} \hat{P}_{\text{on,dyn}}(\text{CC}_{j,k}), \tag{78a}$$

$$\text{subject to} \qquad \text{Constraints (51b)-(51f)}, \tag{78b}$$

which, after similar transformations as the one followed during the analysis of the energy minimization problem, becomes equivalent to

$$\text{minimize} \qquad \ln\left(\sum_{i} \sum_{j} \sum_{k} \exp\left(\ln(\upsilon) + \ln\left(w_{j,k}\frac{\eta_{i,j,k}}{h_{i,j,k}}\right) + \frac{1}{\beta}\frac{\xi_{i,j,k} r_i}{\rho_{j,k} y_{i,j,k}} \ln(2)\right)\right), \tag{79a}$$

$$\text{subject to} \qquad \text{Constraints (51b)-(51f)}. \tag{79b}$$

Applying the log-sum-exp approximation described in Section 3.3.1, the above optimization problem can be approximated as

$$\begin{aligned} \text{minimize} \quad & \max_{i,j,k} \left[\ln(\upsilon) + \ln\left(w_{j,k} \frac{\eta_{i,j,k}}{h_{i,j,k}} \right) + \frac{1}{\beta} \frac{\xi_{i,j,k} r_i}{\rho_{j,k} y_{i,j,k}} \ln(2) \right], && \text{(80a)} \\ \text{subject to} \quad & \text{Constraints (51b)-(51f).} && \text{(80b)} \end{aligned}$$

For $\upsilon = 1$, the above problem is reduced to the original energy minimization problem. Therefore, for $\upsilon = 1$ and following a similar development as in Section 3.3.1, the optimization problem (80) is equivalent to a single convex feasibility problem

$$\begin{aligned} \text{find} \quad & \xi, y, && \text{(81a)} \\ \text{subject to} \quad & \xi_{i,j,k} r_i \ln(2) - \beta \rho_{j,k} y_{i,j,k} \left[(\chi^* - \ln(\upsilon)) - \ln\left(w_{j,k} \frac{\eta_{i,j,k}}{h_{i,j,k}} \right) \right] \leq 0, \quad \forall i, j, k, && \text{(81b)} \\ & \text{Constraints (51b)-(51f).} && \text{(81c)} \end{aligned}$$

If we consider $0 < \upsilon < 1$ in the above problem, we note that its impact translates into increasing the effective threshold $(\chi^* - \ln(\upsilon))$ of the optimization problem. More importantly, in the above feasibility problem, the factor υ is part of the constraint rather than of the objective function. Therefore, the problem above does not suffer from υ not impacting the optimization problem (77), as was the case when we directly used problem (74). So, combining the above problem with that of the capacity maximization, we obtain that the scalarized multi-objective optimization is

$$\begin{aligned} \text{minimize} \quad & -f_3(\xi), && \text{(82a)} \\ \text{subject to} \quad & \xi_{i,j,k} r_i \ln(2) - \beta \rho_{j,k} y_{i,j,k} \left[(\chi^* - \ln(\upsilon)) - \ln\left(w_{j,k} \frac{\eta_{i,j,k}}{h_{i,j,k}} \right) \right] \leq 0, \quad \forall i, j, k, && \text{(82b)} \\ & \text{Constraints (74b)-(74f).} && \text{(82c)} \end{aligned}$$

From the above problem formulation, we can also directly obtain the optimum UE-to-CC association policy without the need to find the solution. As in the case of the energy minimization problem, there are two possible necessary conditions for constraint (82b) to be satisfied:

$$(\chi^* - \ln(\upsilon)) - \ln\left(w_{j,k}\frac{\eta_{i,j,k}}{h_{i,j,k}}\right) > 0, \tag{83}$$

or

$$\xi_{i,j,k} = y_{i,j,k} = 0. \tag{84}$$

Therefore, if

$$(\chi^* - \ln(\upsilon)) \leq \ln\left(w_{j,k}\frac{\eta_{i,j,k}}{h_{i,j,k}}\right), \tag{85}$$

then

$$\xi_{i,j,k} = y_{i,j,k} = 0. \tag{86}$$

Equivalently, we can state that the optimum UE-to-CC association policy for any given υ in the multi-objective optimization of energy minimization and capacity maximization is for UE i to associate with $\mathrm{CC}_{j,k}$ if and only if

$$\frac{h_{i,j,k}}{\eta_{i,j,k}} > w_{j,k}\exp\Big(-\big(\chi^* - \ln(\upsilon)\big)\Big). \tag{87}$$

Once the UEs associate with the CCs, the values of the optimization variables ξ and y will

depend on the particular function f_3 utilized as objective function of the capacity maximization problem. It is important to note that, to perform the multi-objective optimization, we need to find the value of χ^* only once, and then the UE-to-CC association policy is defined by that value, the operator-defined υ, and the specific capacity maximization function of interest.

3.4 Performance Evaluation

In this section, we evaluate the performance of our proposed algorithms for MSCA-enabled HetNets to minimize the energy consumption and balance it with the capacity maximization. The simulation parameters are shown in Table 3. Based on the bandwidth parameters, we have that layer 1 (L1), layer 2 (L2), and layer 3 (L3) provide 18.6%, 46.51%, and 34.88% of the network capacity, respectively. Thus, L1 is meant to provide basic coverage, L2 is meant to provide basic capacity, and L3 is meant to enhance the capacity. BSs per layer and active UEs are uniformly distributed across the total coverage area.

Table 3: Simulation parameters for multi-layer HetNets with MSCA.

Parameter	**Value**
Bandwidth of a CC (per layer)	[20, 10, 2.5] MHz
Max. spectral efficiency	4.8 bps/Hz
Total coverage area	1km x 1km
Number of active UEs	50
Altitude of UEs	1.5 m
Number of layers	3
Type of BSs (per layer)	[macro,pico,pico]
Number of BSs (per layer)	[1,5,15]
Altitude of BSs (per layer)	[25,20,10]m
Power weight of a CC (per layer)	[2.66,3.1,4.0]

For the path loss, we use the following 3GPP models for heterogeneous networks in outdoor scenarios (distance D (in km))[40]:

$$\text{(macro)PL} = 128.1 + 37.6\log(\text{D}), \tag{88}$$

$$\text{(pico)PL} = 140.7 + 36.7\log(\text{D}). \tag{89}$$

To evaluate the overall performance of the energy-saving algorithm, we applied it to 100 different scenarios generated using the parameters from Table 3. For each scenario, Algorithm 4 was evaluated for a minimum QoS varying in the range [1, 10]Mbps. To solve the convex feasibility problem in step 4 of the algorithm, we utilized CVX, a Matlab-based modeling system for convex optimization, together with MOSEK, one of the leading commercial software products for large-scale optimization problems. In 91% of the 1000 scenario-QoS combinations and a lower bound $\epsilon = 0.01$, our algorithm required 11 iterations to converge to a solution and 12 iterations in the rest of the cases. This result highlights the high convergence rate achieved by the initial estimation of the interval $[l_1, l_2]$ from Eq. (66) and Eq. (67).

Figure 17 depicts the percent of UEs using MSCA and the mean UE spectral efficiency. From this figure, we observe that the percent of UEs using MSCA increases from less than 5% to nearly 35% as the minimum QoS requirement increases from 1 to 10Mbps. For this same variation of the minimum QoS requirement, the mean UE spectral efficiency grows from nearly 2.5bps/Hz to 4.7bps/Hz, i.e., it almost reaches the maximum spectral efficiency of 4.8bps/Hz.

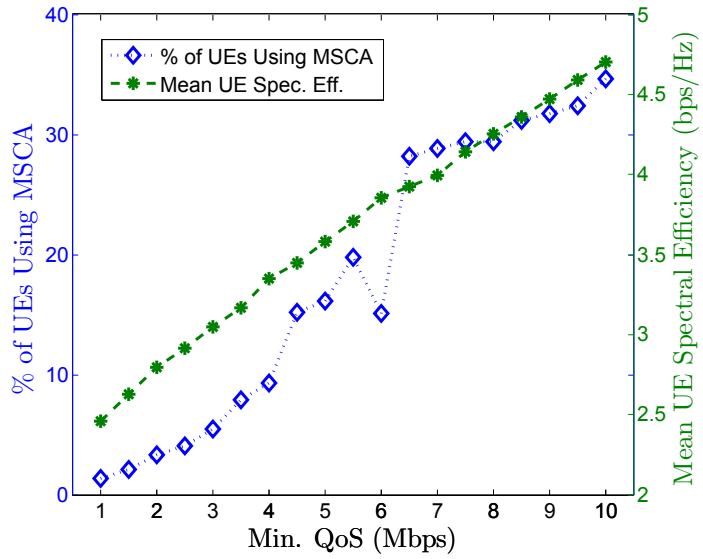
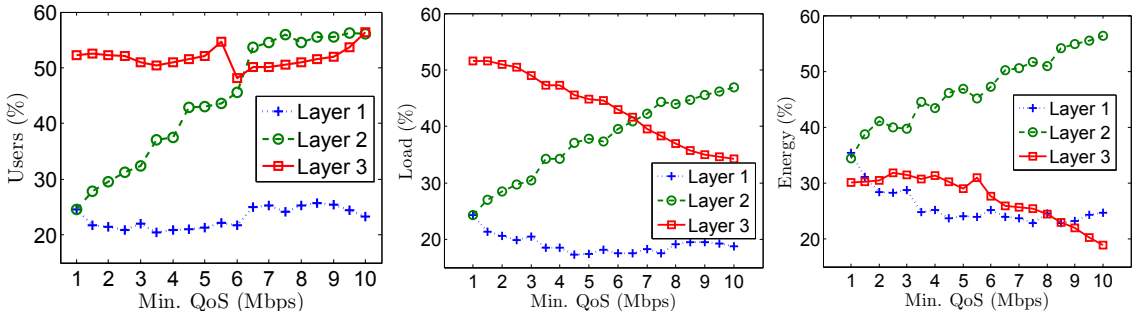


Figure 17: Percent of UEs using MSCA and mean UE spectral efficiency.

The fact that most UEs are operating at nearly maximum spectral efficiency prevents more UEs from applying MSCA since such event would require a first set of UEs to empty part of its currently allocated spectrum so that a second set of UEs, currently connected to other layers, can utilize the freed spectrum. However, such release of spectrum would imply that the UEs of the first set have to further increase their own spectral efficiency.

Figure 18 shows several per-layer metrics. In Figure 18a, we depict how the UEs associate with each BS layer. Here, we observe that as the minimum QoS requirement increases, the percent of UEs associated with L1 and L3 experiences small variations, indicating that most UEs remain connected to those layers. However, the percent of UEs attached to L2, the layer with highest capacity, increases significantly, indicating that most UEs are applying MSCA by connecting to an additional CC in L2.



(a) Users per layer. (b) Load per layer. (c) Energy per layer.

Figure 18: Users, load, and energy per layer.

In Figure 18b, we observe that as the minimum QoS requirement increases, the percent of the load carried by L3 decreases from 51% to 34% while that of L2 increases from 24% to 47%. From Figure 18c, we observe that the change in the load managed by L2 and L3 produces a nearly equivalent change in the percent of energy consumption. That of L3 decreases from 30% to 19% while that of L2 increases from 35% to 56%.

An additional metric of interest is the value of χ^* as the minimum QoS requirement increases. This behavior is depicted in Figure 19.

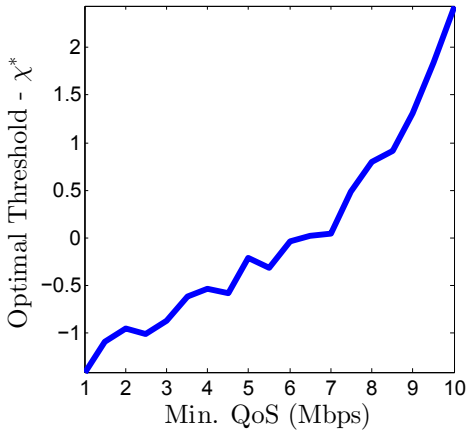


Figure 19: χ^* vs. minimum QoS requirement.

When the minimum QoS requirement is less than 7Mbps, χ^* increases almost linearly from -1.4 to 0.05. However, beyond 7Mbps, χ^* increases rapidly until reaching a value of 2.44.

To quantify the amount of energy savings provided by our energy-saving algorithm, as well as to characterize the energy-capacity trade-off in an MSCA-enabled HetNet, we take a single instance of a HetNet generated with the parameters of Table 3, and analyze its performance as the factor υ varies from 0 to 1, representing a shift from the capacity maximization to the energy minimization objective. We analyze the balance of energy minimization against three objective functions for the capacity maximization:

- Classical capacity maximization:

$$f_{3,1}(\xi) = \sum_i \sum_j \sum_k r_i \xi_{i,j,k}. \tag{90}$$

- Global UE fairness:

$$f_{3,2}(\xi) = \sum_i \log\left(\sum_j \sum_k r_i \xi_{i,j,k}\right). \tag{91}$$

- Per-BS UE fairness: b.

$$f_{3,3}(\xi) = \sum_j \sum_i \log\left(\sum_k r_i \xi_{i,j,k}\right). \tag{92}$$

In Figure 20, we show the percent of UEs that use MSCA. For the classical capacity maximization objective, we observe that very few UEs apply MSCA. This behavior occurs because such objective tends to favor UE-to-CC links with high SINR; therefore, MSCA links with distant BSs tend to be disregarded. On the other extreme, we have the per-BS UE fairness. In this case, the number of UEs applying MSCA is over 90%. This behavior occurs because each BS tries to provide a fair amount of throughput to all the UEs that it can potentially serve; therefore, this objective function encourages the application of MSCA among all the UEs that are under the coverage of more than one layer. We observe that the global UE fairness, with the use of MSCA decreasing from 64% to 30% as υ varies from 0 to 1, falls roughly in the middle between the other two extremes - capacity and

per-BS UE fairness.

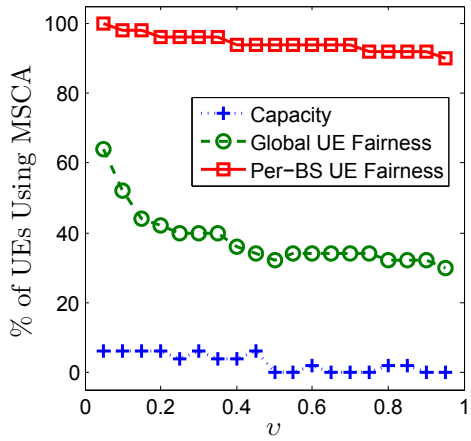
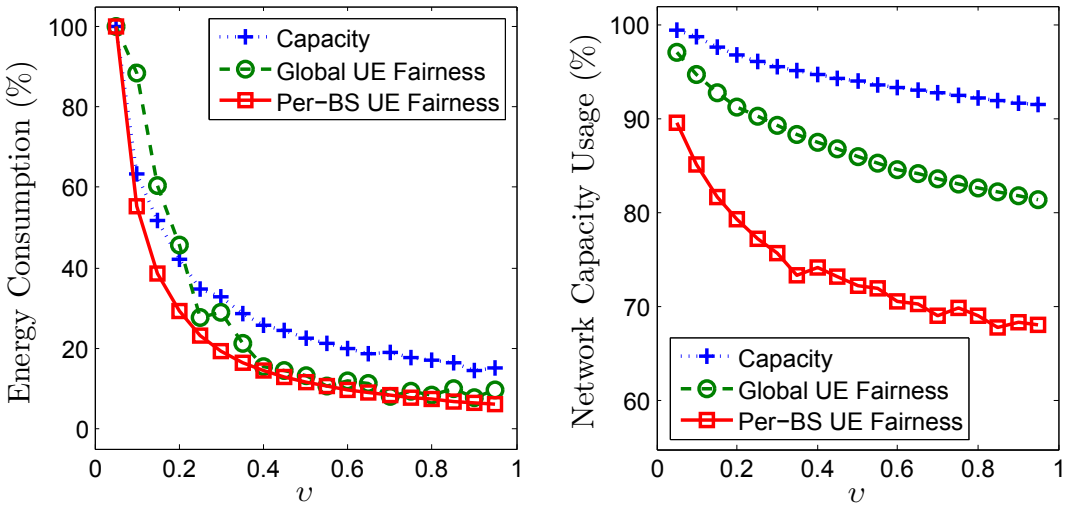


Figure 20: MSCA UEs in the energy-capacity optimization.

In Figure 21, we depict the capacity usage and the energy consumption. From Figure 21a, we observe that by applying the energy minimization algorithm it is possible to decrease the energy consumption to at least 15% of its maximum for all the capacity objectives.



(a) Energy consumption.

(b) Network capacity usage.

Figure 21: Energy consumption and capacity usage in the energy-capacity optimization.

The effect of υ on the network capacity usage is shown in Figure 21b. From this graph, we observe that minimizing the energy consumption has the greatest impact, from 90% to

68%, on the capacity usage for the per-BS UE fairness objective. Conversely, the classical capacity objective experiences the least impact, from 99% to 91.45%.

From the above graphs, we can now generate the energy-capacity trade-off curve for the MSCA-enabled HetNet, as shown in Figure 22. From this graph, we observe that reducing the capacity usage by as little as 5% allows to significantly increase the energy savings in all of the three capacity objectives. However, even though it is possible to augment the energy savings by further reducing the capacity usage, the return from such reduction tends to diminish, particularly for the per-BS UE fairness objective.

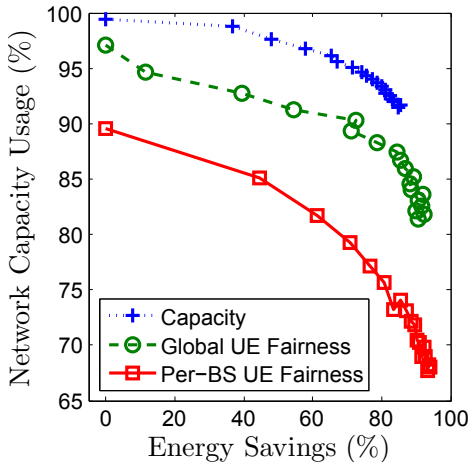


Figure 22: Trade-off curve for the energy savings vs. capacity usage.

3.5 Conclusions

MSCA has been introduced in LTE-Advanced as a mechanism to increase the amount of bandwidth available to the users in HetNets. However, existing work has focused on exploiting the use of MSCA to maximize the network capacity, disregarding the energy efficiency of MSCA. In this chapter, we studied the problem of minimizing the energy consumption in MSCA-enabled HetNets and developed an efficient algorithm to solve it. We showed that, by utilizing a quasi-convex relaxation, we are able to not only solve the problem, but also to establish a clear and simple cell-association policy. Moreover, we showed how this cell-association policy can be easily adjusted to obtain a new policy that balances

the conflicting objectives of energy minimization and capacity maximization. Through extensive simulations, we characterized the effects of our algorithm on the percent of load, users, and energy per layer, as well as on the percent of UEs that use MSCA and their average spectral efficiency. In addition, we obtained the trade-off curve between the energy minimization and capacity maximization and found that a large amount of energy savings can be achieved in an MSCA-enabled HetNet by reducing the network capacity usage by as little as 5%.

CHAPTER 4

REDUCING THE USER EQUIPMENT ENERGY CONSUMPTION THROUGH CROSS-CARRIER-AWARE DISCONTINUOUS RECEPTION

To reduce the energy consumption at the UE, 3GPP introduced in LTE the concept of discontinuous reception (DRX). Nevertheless, existing models for the LTE DRX and their extension to scenarios that support carrier aggregation have several drawbacks. In this chapter, we utilize a semi-Markov Chain to model the operation of the LTE DRX and characterize its performance metrics. Then, we exploit the new features introduced in LTE-A to develop a novel cross-carrier-aware DRX for scenarios that support carrier aggregation. Our DRX solution is modeled and examined in detail, and analytical expressions for its performance metrics are obtained. The accuracy of our modeling approach, for both the classical and our novel LTE DRX, is validated through extensive simulations. We then evaluate the performance of our cross-carrier-aware DRX and demonstrate that it significantly outperforms the classical DRX, particularly under the most challenging conditions of low tolerable delay.

4.1 Motivation and Related Work

In addition to improving the energy efficiency of the hardware components in the UE, the main technique to reduce the energy consumption at the UE is the use of discontinuous reception (DRX) [70]. Initially introduced by 3GPP in UMTS and later in LTE, DRX holds as its main objective the possibility for the UE to turn off most of its circuitry during the periods of time when the traffic is not being exchanged with the BS. By doing so, the energy consumption of the UE can be brought to a minimum level during the "sleep" periods and, therefore, the on-board energy utilization of the UE can be maximized. The BS, always aware of the DRX state of the UE, buffers any packet received while the UE is "sleeping" and sends it once the UE "awakes". The buffered packets experience additional delay,

which can be extremely detrimental, particularly for delay-sensitive applications. Therefore, it is essential to choose the optimal values for the DRX parameters that maximize the energy savings without sacrificing the delay metrics. These values can only be optimally selected when there is a clear understanding of their impact on the DRX performance.

The performance of the LTE DRX has been previously studied in the literature, but with many simplifying assumptions that severely hinder the applicability of the results.

1. It is commonly assumed that the packet arrivals, departures, the service of a packet, and many other events can occur at any time [71][23][72]. While this assumption allows for simplified formulations, it is not valid for all events. In particular, in LTE, when a packet arrives during a given subframe, it cannot be scheduled or sent during that same subframe since the scheduling grant for that subframe was previously sent during its first one to three OFDM symbols [73][74]. Therefore, such packet must wait at least until the start of the next subframe to be scheduled. Thus, the previously described assumption leads to

    - the underestimation of the time needed to empty the buffer,
    - the underestimation of the time spent waiting for a scheduling grant from the BS, and
    - the overestimation of the energy savings.

2. It is commonly assumed that the packet inter-arrival and service times follow an exponential distribution [71][23][72][75][76][77][78][79][80]. As before, such assumption allows simplified formulations, but cannot be readily extended to cases where different probability distributions appear.

3. Some DRX parameters, such as the *OnDurationTimer*, have not been taken into account to simplify the formulation [71][23][72].

4. The existing literature has focused on analyzing the delay experienced by packets

that arrive at the BS while the UE is "sleeping" and disregarded the effect that the buffered packets have on the ones that arrive after the "sleep" period ends. This leads to an incomplete characterization of the DRX impact on the packet delay.

Moreover, as discussed in Chapter 3, with the introduction of CA and MSCA, an LTE-A UE can simultaneously utilize up to five CCs, each one of up to 20MHz, to communicate with a BS. Compared to LTE, such CA increases the energy consumption as much as five times. At the LTE-A UE, the use of DRX still remains the most viable option to reduce the energy consumption. However, the existing work on LTE-A DRX has focused on using either the same DRX parameters for all CCs, or completely different parameters for each CC [75][81][70]. The first approach is simple, but extremely rigid and inefficient, since all the CCs are active, regardless of traffic being transmitted. The second approach provides flexibility per CC, but completely discards the cross-carrier awareness that exists at the BS and the UE.

In this chapter, our objective is to address the aforementioned limitations in LTE-A DRX by introducing a novel mechanism of cross-carrier-aware DRX. We exploit the new features introduced in LTE-A, namely cross-carrier resource assignment and signaling-reduced CCs. With our proposed solution, we significantly decrease the energy consumption across all CCs by allowing them to enter into a "deep sleep" mode, selectively triggering their reactivation, and supporting per-CC DRX parameters. At the same time, we account for the current 3GPP specifications regarding the use of cross-carrier resource assignment. To develop our cross-carrier-aware DRX, we first design and analyze accurate models that address the limitations of existing modeling approaches of the classical DRX and its achievable energy savings. We then utilize these models as the basis for our cross-carrier-aware DRX.

The rest of this chapter is organized as follows. In Section 4.2, we introduce and analyze the operation of the LTE DRX. In particular, in Section 4.2.1, we describe a semi-Markov Chain model to analyze the LTE DRX. The stationary probability and the holding time of

the states in such model are analyzed in Sections 4.2.2 and 4.2.3, respectively. The performance metrics are then characterized in Section 4.2.4 and evaluated in Section 4.2.5. Then, we introduce our cross-carrier-aware DRX in Section 4.3. In particular, in Section 4.3.1, we present a semi-Markov Chain model to analyze our proposed solution. Then, the stationary probability and holding time of its states are analyzed in Sections 4.3.2-4.3.3 and 4.3.4, respectively. In Section 4.3.5, we characterize the performance metrics of our proposed DRX scheme and, in Section 4.3.6, compare them to the ones of the LTE DRX. Finally, we present the conclusions in Section 4.4.

4.2 DRX Analysis

The basic operation of DRX in LTE is shown in Figure 23. While the BS is actively sending packets to the UE, the latter is in a continuous reception state. When the packet transmission from the BS is interrupted, the UE remains waiting for additional packets for a duration controlled by the *drx-InactivityTimer*, as defined by 3GPP terminology. During this inactivity period, the UE actively monitors the physical downlink control channel (PDCCH) from the BS, looking for a scheduling grant. If the UE receives such grant before the *drx-InactivityTimer* expires, it returns to the continuous reception state. Otherwise, the UE starts a short DRX cycle. Such cycle is divided into two parts: an "on" period controlled by the *onDurationTimer*, and a "sleep" period. If the UE receives a scheduling grant from the BS before the *onDurationTimer* expires, it returns to the continuous reception state. Otherwise, it transitions to the "sleep" state, where it can turn off most of its circuitry to reduce its energy consumption. The BS is always aware of the DRX parameters of the UEs; therefore, if any packet intended for a given UE arrives while that UE is in a "sleep" state, then the BS buffers the packet and sends it to the UE once the latter exits the "sleep" state. If no packets are awaiting at the BS for the UE by the time it exits the "sleep" state, the UE starts a new short DRX cycle. As long as the UE receives no scheduling grant, it repeats up to N short DRX cycles. If by the end of the N-th cycle the UE has not received a scheduling

grant, it transitions to a long DRX cycle. The only difference between the two cycles is that the duration of the "sleep" period is greater in the long DRX cycle; therefore, the long DRX cycle further reduces the energy consumption. The UE repeats the long DRX cycle until a scheduling grant arrives from the BS and triggers a transition to the continuous reception state.

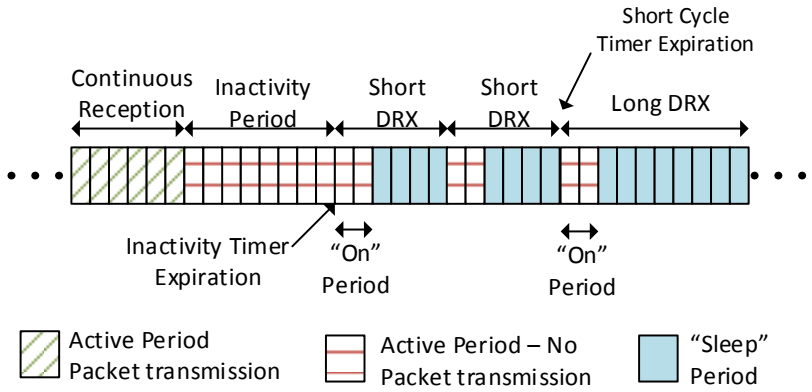


Figure 23: LTE DRX operation.

The DRX operation previously described is affected by the subframe-based[13] transmissions in LTE. Here, a packet that arrives to the BS during a given subframe x cannot be sent immediately to the UE; instead, it must wait at least for the next subframe $x + 1$ to be scheduled by the BS. This constraint has two important effects:

1. Any packet that arrives to the BS during the subframe where the UE was executing its last subframe of the "on" period cannot be sent to the UE until the latter "wakes up" from the "sleep" period.

2. Compared to a continuous-time model, where the continuous reception state for a given UE finishes as soon as the BS has no more packets for that UE (regardless of the subframe alignment), in LTE/LTE-A, the continuous reception state of that UE can only finish at the end of subframe x if (a) by the end of subframe x the BS has sent the UE all the packets present in the buffer at the end of subframe $x - 1$, and (b) no new packets arrive during subframe x.

[13] In LTE, the length of a subframe is typically 1ms.

We account for these two constraints by considering a discrete-time system with a late-arrival model, as described in Section 4.2.1.

4.2.1 DRX Model

The DRX operation previously described is captured as a finite state machine (FSM) model, as seen in Figure 24. There are three important features to highlight in this FSM model:

1. The amount of time spent in each state varies among them. For the "sleep" period, such time is fixed, while for the rest of the states it is a random variable with a different probability mass function per state.

2. The short DRX cycle is repeated up to N times before the long DRX cycle is reached. Therefore, memory is required to keep track of the number of short DRX cycles that have been previously executed.

3. It is possible to reach the continuous reception state from every other state. However, the amount of time spent in the continuous reception state depends directly on the previously executed state.

For these three reasons, the DRX operation cannot be directly modeled as a typical Markov Chain and requires an expansion beyond the FSM model.

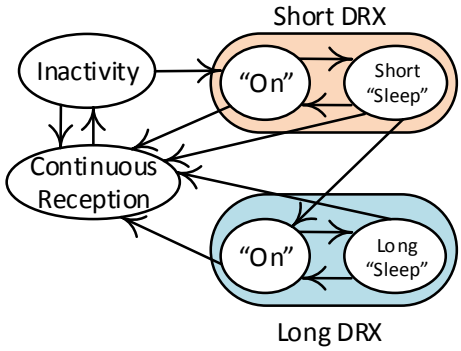


Figure 24: LTE DRX finite state machine.

We utilize a semi-Markov Chain with late arrival to model the DRX operation, as shown in Figure 25. The description of each state is shown in Table 4. Each of the N possible short

DRX cycles is explicitly modeled as a pair of S_{2i-1} and S_{2i} states. The former represents the "on" period of the i-th short DRX cycle, while the latter represents the corresponding "sleep" period. Similarly, the long DRX cycle is modeled by the pair of states G_1 and G_2, representing its "on" and "sleep" periods, respectively. The inactivity period is modeled by the state B.

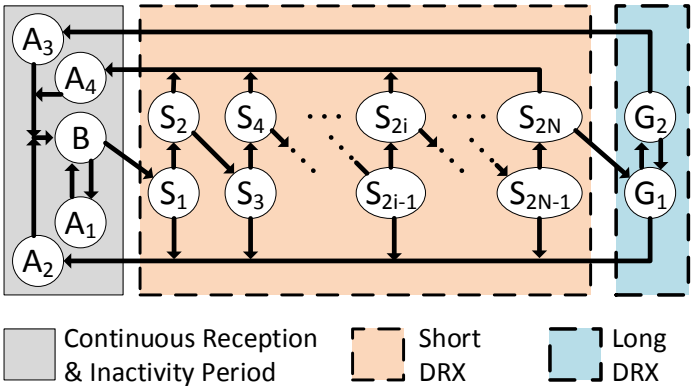


Figure 25: LTE DRX semi-Markov Chain model.

Table 4: LTE DRX states description.

State	Description
B	Inactivity period.
S_{2i}	"Sleep" period of the i-th short DRX cycle.
S_{2i-1}	"On" period of the i-th short DRX cycle.
G_1	"On" period of the long DRX cycle.
G_2	"Sleep" period of the long DRX cycle.
A_1	Continuous reception following state B.
A_2	Continuous reception following states S_{2i-1} and G_1.
A_3	Continuous reception following state G_2.
A_4	Continuous reception following state S_{2i}.

Rather than using a single state to model the continuous reception, we utilize four different states. The reasoning behind this approach is that the number of packets sent by

the BS during continuous reception and, therefore, the amount of time spent in continuous reception depend on the UE state at the moment that the packet triggering the continuous reception arrives at the BS. For example, the expected number of packets received at the BS during the "sleep" period of a short DRX cycle is smaller than that received during the "sleep" period of the long DRX cycle. Therefore, the duration of state A_3, defined in Table 4, is longer than that of A_4.

The transitions in Figure 25 are controlled by the parameters in Table 5 and reflect the DRX operation described in Section 4.2.

Table 5: LTE DRX parameters.

Parameter	Description
T_α	Inactivity period length.
T_β	Short DRX cycle length.
T_γ	Long DRX cycle length.
N	Number of short DRX cycles.
T_{on}	"On" period length.

The transitions across states occur as follows:

- If the UE is in the inactivity state B and receives no scheduling grant after T_α subframes, it transitions to the "on" state S_1 of the first short DRX cycle. Otherwise, it transitions to state A_1. A transition to S_1 implies that the BS received no packets during the previous T_α subframes.

- If the UE is in the "on" state $S_{2i-1}, i \in [1, N]$, of the i-th short DRX cycle and receives no scheduling grant after T_{on} subframes, it transitions to the corresponding "sleep" state S_{2i}. Otherwise, it transitions to the continuous reception state A_2. A transition to S_{2i} implies that the BS received no packets during the previous $T_{\text{on}} - 1$ subframes. Given the structure of the late-arrival model, a packet that arrives at the BS during

the last subframe of an S_{2i-1} state of a given UE must wait for the latter to return from its "sleep" state to be scheduled.

- If the UE is in the "sleep" state $S_{2i}, i \in [1, N-1]$, of the i-th short DRX cycle and the BS has no packets for the UE by the time the UE "awakes," the latter transitions to the "on" state S_{2i+1}. Otherwise, it transitions to the continuous reception state A_4. A transition to S_{2i+1} implies that the BS received no packets during the previous $T_\beta - T_{\text{on}} + 1$ subframes. This time accounts for the $T_\beta - T_{\text{on}}$ subframes that the UE spent in the "sleep" state S_{2i} and the last subframe of the "on" state S_{2i-1} that preceded state S_{2i}.

- If the UE is in the "sleep" state S_{2N} of the N-th short DRX cycle and the BS has no packets for the UE by the time the UE "awakes," the latter transitions to the "on" state G_1 of the long DRX. Otherwise, it transitions to the continuous reception state A_4. A transition to G_1 implies that the BS received no packets during the previous $T_\beta - T_{\text{on}} + 1$ subframes.

- If the UE is in the "on" state G_1 of the long DRX cycle and receives no scheduling grant after T_{on} subframes, it transitions to the "sleep" state G_2. Otherwise, it transitions to the continuous reception state A_2. Similarly to the states $S_{2i-1}, i \in [1, N]$, a transition to the "sleep" state implies that the BS received no packets during the previous $T_{\text{on}} - 1$ subframes.

- If the UE is in the "sleep" state G_2 of the long DRX cycle and the BS has no packets for the UE by the time the UE "awakes," the latter transitions back to the "on" state G_1 of the long DRX. Otherwise, it transitions to the continuous reception state A_3. A transition to G_1 implies that the BS received no packets during the previous $T_\gamma - T_{\text{on}} + 1$ subframes. This time accounts for the $T_\gamma - T_{\text{on}}$ subframes that the UE spent in the "sleep" state G_2 and the last subframe of the "on" state G_1 that preceded state G_2.

- If the UE is in a continuous reception state A_i, it will remain in that state until the BS has no more packets left to send to the UE. Once the UE detects, due to the absence of a scheduling grant, that the no-packet event occurs, it transitions to the inactivity state B.

In our model, we made two important considerations:

1. We used A_2 to represent the continuous reception state that follows the "on" state of the short and long DRX cycles. This is possible because the number of packets that are expected to be transmitted during continuous reception if any of these DRX cycles is interrupted during the "on" state is the same. As we show in Section 4.2.3, this quantity corresponds to the expected number of packets received at the BS in one subframe.

2. A_1 and A_2 can be merged into a single state following a similar argument as above. Nevertheless, we have chosen to avoid merging them, so we can better reflect the mapping between the DRX operation and the semi-Markov Chain model.

4.2.2 Stationary Probability - Embedded Markov Chain

In this section, we present the calculation of the stationary probabilities of the embedded Markov Chain (EMC), which are later used in Section 4.2.4 to calculate the performance metrics of the DRX.

Utilizing the notation $p_{U,U'}$ for the transition probability from state U to state U' and π_U for the stationary probability of state U in the EMC, the stationary probabilities π of the EMC follow these relationships:

$$\pi_B = \sum_{i=1}^{4} \pi_{A_i}, \tag{93a}$$

$$\pi_{S_1} = \pi_B p_{B,S_1}, \tag{93b}$$

$$\pi_{S_i} = \pi_{S_{i-1}} p_{S_{i-1},S_i}, \qquad i \in [2, 2N], \tag{93c}$$

$$\pi_{G_1} = \pi_{S_{2N}} p_{S_{2N},G_1} + \pi_{G_2} p_{G_2,G_1}, \tag{93d}$$

$$\pi_{G_2} = \pi_{G_1} p_{G_1,G_2}, \tag{93e}$$

$$\pi_{A_1} = \pi_B p_{B,A_1}, \tag{93f}$$

$$\pi_{A_2} = \sum_{i=1}^{N} \pi_{S_{2i-1}} p_{S_{2i-1},A_2} + \pi_{G_1} p_{G_1,A_2}, \tag{93g}$$

$$\pi_{A_3} = \pi_{G_2} p_{G_2,A_3}, \tag{93h}$$

$$\pi_{A_4} = \sum_{i=1}^{N} \pi_{S_{2i}} p_{S_{2i},A_4}. \tag{93i}$$

From these expressions we obtain

$$\pi_B = \frac{\pi_{S_1}}{p_{B,S_1}}, \tag{94a}$$

$$\pi_{S_i} = \pi_{S_1} \prod_{j=2}^{i} p_{S_{j-1},S_j}, \qquad i \in [2, 2N], \tag{94b}$$

$$\pi_{G_1} = \pi_{S_1} \omega\theta, \tag{94c}$$

$$\pi_{G_2} = \pi_{S_1} \omega\theta p_{G_1,G_2}, \tag{94d}$$

$$\pi_{A_1} = \pi_{S_1} \frac{p_{B,A_1}}{p_{B,S_1}}, \tag{94e}$$

$$\pi_{A_2} = \pi_{S_1} \left[p_{S_1,A_2} + \omega\theta p_{G_1,A_2} + \sum_{i=2}^{N} \left(p_{S_{2i-1},A_2} \prod_{j=2}^{2i-1} p_{S_{j-1},S_j} \right) \right], \tag{94f}$$

$$\pi_{A_3} = \pi_{S_1} \omega\theta p_{G_1,G_2} p_{G_2,A_3}, \tag{94g}$$

$$\pi_{A_4} = \pi_{S_1} \sum_{i=1}^{N} \left(p_{S_{2i},A_4} \prod_{j=2}^{2i} p_{S_{j-1},S_j} \right), \tag{94h}$$

where

$$\omega = p_{S_{2N},G_1} \prod_{j=2}^{2N} p_{S_{j-1},S_j}, \qquad \theta = (1 - p_{G_1,G_2} p_{G_2,G_1})^{-1}. \tag{95}$$

With these expressions, the value of π_{S_1} becomes

$$\pi_{S_1} = \left[1 + \frac{2}{p_{B,S_1}} + \omega\theta(1 + p_{G_1,G_2}) + \sum_{i=2}^{2N} \prod_{j=2}^{i} p_{S_{j-1},S_j} \right]^{-1}, \tag{96}$$

which can be plugged in Eq. (94) to obtain the stationary probabilities of all other states of the EMC model.

To find the actual values of the above expressions, it is necessary to obtain the transition probabilities, which depend on the BS packet arrival model, probability distribution, and the service discipline. For the latter, we consider that the BS has a separate and infinite buffer for each UE, as is regularly done in the existing literature. Therefore, the rest of the analysis focuses on the DRX operation for a single UE.

For the packet arrival model, we utilize a late-arrival model [82]. In such model, packets are considered to arrive just before the end of a subframe. If a given packet arrives during subframe x, and no other packets are in the BS by the end of x, then the service of the packet is immediately started in subframe $x + 1$. The time between the packet arrival instant and the start of the next subframe is not included in the calculation of the packet waiting time. For the packet arrival distribution, we consider that the number of packets that arrive at the BS during successive subframes constitutes a sequence of independent and identically distributed (i.i.d.) random variables. We use Λ to denote the number of packets that arrive in a single subframe. The probability mass function (PMF) of Λ is defined as

$$\lambda(k) \triangleq \text{Prob}\{\Lambda = k\}, \qquad k = 0, 1, 2..., \tag{97}$$

and its mean value is defined as

$$\lambda \triangleq E\{\Lambda\}, \tag{98}$$

where $E\{\Lambda\}$ denotes the expected value of Λ. We denote as X the service time (measured

in subframes) of a single packet. The PMF of X is defined as

$$b(l) = \text{Prob}\{X = l\}, \qquad l = 1, 2, ..., \tag{99}$$

and its mean value is defined as

$$b \triangleq E\{X\}. \tag{100}$$

Having the arrival model, the service discipline, and the PMF of the packet arrivals and the service time, we proceed to characterize the transition probabilities. To achieve this, we (a) apply the conditions that trigger each transition, as described in Section 4.2.1, and (b) consider that the probability of a BS receiving no packets in a time period of length v is equal to $[\lambda(0)]^v$ since the number of packet arrivals in successive subframes constitutes a sequence of i.i.d. random variables. Consequently, the probability that the BS receives at least one packet in a time period of length v is equal to $1 - [\lambda(0)]^v$.

$$p_{B,S_1} = [\lambda(0)]^{T_\alpha}, \tag{101a}$$

$$p_{B,A_1} = 1 - [\lambda(0)]^{T_\alpha}, \tag{101b}$$

$$p_{S_{2i-1},S_{2i}} = [\lambda(0)]^{T_{\text{on}}-1}, \qquad i \in [1, N], \tag{101c}$$

$$p_{S_{2i-1},A_2} = 1 - [\lambda(0)]^{T_{\text{on}}-1}, \qquad i \in [1, N], \tag{101d}$$

$$p_{S_{2i},S_{2i+1}} = [\lambda(0)]^{T_\beta - T_{\text{on}}+1}, \qquad i \in [1, N-1], \tag{101e}$$

$$p_{S_{2i},A_4} = 1 - [\lambda(0)]^{T_\beta - T_{\text{on}}+1}, \qquad i \in [1, N], \tag{101f}$$

$$p_{S_{2N},G_1} = [\lambda(0)]^{T_\beta - T_{\text{on}}+1}, \tag{101g}$$

$$p_{G_1,G_2} = [\lambda(0)]^{T_{\text{on}}-1}, \tag{101h}$$

$$p_{G_1,A_2} = 1 - [\lambda(0)]^{T_{\text{on}}-1}, \tag{101i}$$

$$p_{G_2,G_1} = [\lambda(0)]^{T_\gamma - T_{\text{on}}+1}, \tag{101j}$$

$$p_{G_2,A_3} = 1 - [\lambda(0)]^{T_\gamma - T_{\text{on}}+1}. \tag{101k}$$

Having the transition probabilities, we plug them into Eq. (95) and Eq. (96) and obtain

$$\omega = [\lambda(0)]^{NT_\beta}, \qquad \theta = \left(1 - [\lambda(0)]^{T_\gamma}\right)^{-1}, \tag{102}$$

$$\pi_{S_1} = \left[\frac{2}{[\lambda(0)]^{T_\alpha}} + \left(1 + [\lambda(0)]^{T_{\text{on}}-1}\right)\left(\frac{1 - [\lambda(0)]^{NT_\beta}}{1 - [\lambda(0)]^{T_\beta}} + \frac{[\lambda(0)]^{NT_\beta}}{1 - [\lambda(0)]^{T_\gamma}}\right)\right]^{-1}. \tag{103}$$

By denoting $\phi \triangleq \pi_{S_1}$, the expressions for the stationary probability for the rest of the states become

$$\pi_B = \frac{\phi}{[\lambda(0)]^{T_\alpha}}, \tag{104a}$$

$$\pi_{S_i} = \phi \begin{cases} [\lambda(0)]^{T_\beta(i-1)/2} & : i \text{ is odd} \\ [\lambda(0)]^{T_\beta(i-2)/2}\,[\lambda(0)]^{T_{\text{on}}-1} & : i \text{ is even} \end{cases}, i \in [1, 2N], \tag{104b}$$

$$\pi_{G_1} = \phi \frac{[\lambda(0)]^{NT_\beta}}{1 - [\lambda(0)]^{T_\gamma}}, \tag{104c}$$

$$\pi_{G_2} = \phi \frac{[\lambda(0)]^{NT_\beta}}{1 - [\lambda(0)]^{T_\gamma}}\,[\lambda(0)]^{T_{\text{on}}-1}, \tag{104d}$$

$$\pi_{A_1} = \phi\left(\frac{1}{[\lambda(0)]^{T_\alpha}} - 1\right), \tag{104e}$$

$$\pi_{A_2} = \phi\left(1 - [\lambda(0)]^{T_{\text{on}}-1}\right)\left(\frac{1 - [\lambda(0)]^{NT_\beta}}{1 - [\lambda(0)]^{T_\beta}} + \frac{[\lambda(0)]^{NT_\beta}}{1 - [\lambda(0)]^{T_\gamma}}\right), \tag{104f}$$

$$\pi_{A_3} = \phi\,[\lambda(0)]^{NT_\beta}\left[1 - \frac{1 - [\lambda(0)]^{T_{\text{on}}-1}}{1 - [\lambda(0)]^{T_\gamma}}\right], \tag{104g}$$

$$\pi_{A_4} = \phi\left[1 - [\lambda(0)]^{NT_\beta}\right]\left[1 - \frac{1 - [\lambda(0)]^{T_{\text{on}}-1}}{1 - [\lambda(0)]^{T_\beta}}\right]. \tag{104h}$$

In addition to the stationary probabilities of the EMC, the mean amount of time spent in each state is required to compute the energy savings. In Section 4.2.3, we describe how this amount is obtained.

4.2.3 Holding Time

The holding time H_U of a state U represents the mean amount of time spent in such state. The states corresponding to "sleep" periods have a deterministic holding time, since the

length of the "sleep" periods is a constant:

$$H_{S_{2i}} = T_\beta - T_{\text{on}}, \qquad i \in [1, N], \tag{105a}$$

$$H_{G_2} = T_\gamma - T_{\text{on}}. \tag{105b}$$

The holding time of the "on" periods and the inactivity period can be calculated as follows. Consider the maximum length of any such period to be equal to v, and the amount of time spent in it to be L. Then, the PMF of L is

$$\text{Prob}\{L = i|v\} = \begin{cases} [\lambda(0)]^{i-1}\,(1-\lambda(0)) & : 0 < i < v \\ [\lambda(0)]^{v-1} & : i = v \end{cases}, \tag{106}$$

i.e., $L = i, i < v$, if no packets are received in the first $i-1$ subframes and at least one packet arrives during the i-th subframe. $L = v$ if no packets arrived during the previous $v-1$ subframes, regardless of whether a packet arrives in the last subframe or not. Taking the expected value of L, we obtain:

$$H = E\{L|v\} = \frac{1-[\lambda(0)]^v}{1-\lambda(0)}. \tag{107}$$

For the "on" periods, $v = T_{\text{on}}$ and for the inactivity period, $v = T_\alpha$. Hence,

$$H_{S_{2i-1}} = E\{L|v = T_{\text{on}}\} = \frac{1-[\lambda(0)]^{T_{\text{on}}}}{1-\lambda(0)}, \qquad i \in [1, N], \tag{108a}$$

$$H_{G_1} = E\{L|v = T_{\text{on}}\} = \frac{1-[\lambda(0)]^{T_{\text{on}}}}{1-\lambda(0)}, \tag{108b}$$

$$H_B = E\{L|v = T_\alpha\} = \frac{1-[\lambda(0)]^{T_\alpha}}{1-\lambda(0)}. \tag{108c}$$

States A_1 through A_4 correspond to a busy period in queuing theory terminology [82], from which it follows that if a busy period A starts with R_A packets in the buffer, its duration

L has a mean value of

$$E\{L|R_A\} = R_A \frac{b}{1-\rho}, \tag{109}$$

where b is the expected value $E\{X\}$ of the packet service time and $\rho = b\lambda$. Therefore, applying the law of total expectation, we obtain that

$$E\{L\} = E\{\{L|R_A\}\} = E\{R_A\}\frac{b}{1-\rho}. \tag{110}$$

Hence, to obtain the holding time of the states A_1, A_2, A_3, and A_4, we need the expected value of the number of packets in the BS buffer when each state starts. This value can be obtained as follows. Consider F to be the number of packets received by the BS during time v, and Q to denote F conditioned on at least one packet arrival. As a result,

$$\text{Prob}\{Q = k|v\} = \begin{cases} \frac{\text{Prob}\{F=k|v\}}{1-\text{Prob}\{F=0|v\}} & : k > 0 \\ 0 & : k = 0 \end{cases}, \tag{111a}$$

$$E\{Q|v\} = \frac{E\{F|v\}}{1-\text{Prob}\{F=0|v\}}. \tag{111b}$$

Since F is the sum of v random variables Λ, $E\{F|v\} = vE\{\Lambda\} = v\lambda$. In addition, $\text{Prob}\{F = 0|v\} = [\lambda(0)]^v$, since it represents the probability of no packet arrival in v subframes. With the previous expressions, we obtain

$$E\{Q|v\} = \lambda\frac{v}{1-[\lambda(0)]^v}. \tag{112}$$

$E\{Q|v\}$ represents the expected number of packets buffered at the BS after v subframes, given that at least one packet arrives. This expression allows us to obtain R_{A_i} for each state A_i since each of them is entered when the BS has received at least one packet during a given

number of previous subframes. Specifically,

$$E\{R_{A_1}\} = E\{Q|v = 1\} = \lambda \frac{1}{1 - [\lambda(0)]}, \tag{113a}$$

$$E\{R_{A_2}\} = E\{Q|v = 1\} = \lambda \frac{1}{1 - [\lambda(0)]}, \tag{113b}$$

$$E\{R_{A_3}\} = E\{Q|v = T_\gamma - T_{\text{on}} + 1\} = \lambda \frac{T_\gamma - T_{\text{on}} + 1}{1 - [\lambda(0)]^{T_\gamma - T_{\text{on}} + 1}}, \tag{113c}$$

$$E\{R_{A_4}\} = E\{Q|v = T_\beta - T_{\text{on}} + 1\} = \lambda \frac{T_\beta - T_{\text{on}} + 1}{1 - [\lambda(0)]^{T_\beta - T_{\text{on}} + 1}}. \tag{113d}$$

Hence, the holding times of A_1, A_2, A_3, and A_4 are, respectively,

$$H_{A_1} = E\{R_{A_1}\}\frac{b}{1-\rho} = \frac{\rho}{1-\rho}\frac{1}{1 - [\lambda(0)]}, \tag{114a}$$

$$H_{A_2} = E\{R_{A_2}\}\frac{b}{1-\rho} = \frac{\rho}{1-\rho}\frac{1}{1 - [\lambda(0)]}, \tag{114b}$$

$$H_{A_3} = E\{R_{A_3}\}\frac{b}{1-\rho} = \frac{\rho}{1-\rho}\frac{T_\gamma - T_{\text{on}} + 1}{1 - [\lambda(0)]^{T_\gamma - T_{\text{on}} + 1}}, \tag{114c}$$

$$H_{A_4} = E\{R_{A_4}\}\frac{b}{1-\rho} = \frac{\rho}{1-\rho}\frac{T_\beta - T_{\text{on}} + 1}{1 - [\lambda(0)]^{T_\beta - T_{\text{on}} + 1}}. \tag{114d}$$

Having the holding time of each state, as described in Eq. (105), (108), and (114), we proceed to obtain the energy-savings and delay metrics in Sections 4.2.4.1 and 4.2.4.2, respectively.

4.2.4 Performance Metrics

The two main performance metrics associated with DRX are the amount of energy saved and the packet delay, also known as waiting time in queuing theory terminology.

4.2.4.1 Energy Savings

The amount of energy saved is defined as the amount of time spent in the "sleep" periods. This value can be obtained from the stationary probabilities of the semi-Markov Chain, which we derive from the stationary probabilities of the EMC, Eq. (104a)-(104h).

For any state U, whose stationary probability in the EMC is π_U and whose holding time

is H_U, the stationary probability $\tilde{\pi}_U$ in the semi-Markov Chain is:

$$\tilde{\pi}_U = \frac{\pi_U H_U}{\sum\limits_{\forall U'} \pi_{U'} H_{U'}}. \tag{115}$$

Then, the energy savings τ_β and τ_γ provided by the short and long DRX, respectively, are

$$\tau_\beta = \sum_{i=1}^{N} \tilde{\pi}_{S_{2i}}, \qquad \tau_\gamma = \tilde{\pi}_{G_2}. \tag{116}$$

Replacing the expressions for the holding time and the stationary probability of the EMC, τ_β and τ_γ become

$$\tau_\beta = \frac{(T_\beta - T_{\text{on}}) \frac{1-[\lambda(0)]^{NT_\beta}}{1-[\lambda(0)]^{T_\beta}} [\lambda(0)]^{T_{\text{on}}-1}}{\Psi}, \tag{117a}$$

$$\tau_\gamma = \frac{(T_\gamma - T_{\text{on}}) \frac{[\lambda(0)]^{NT_\beta}}{1-[\lambda(0)]^{T_\gamma}} [\lambda(0)]^{T_{\text{on}}-1}}{\Psi}, \tag{117b}$$

where

$$\begin{aligned}\Psi = {} & \frac{1}{[\lambda(0)]^{T_\alpha}} \left[H_B + \left(1 - [\lambda(0)]^{T_\alpha}\right) H_{A_1}\right] + [\lambda(0)]^{NT_\beta} H_{A_3} + \left(1 - [\lambda(0)]^{NT_\beta}\right) H_{A_4} \\ & + \frac{1 - [\lambda(0)]^{NT_\beta}}{1 - [\lambda(0)]^{T_\beta}} \left[H_{S_1} + H_{S_2} [\lambda(0)]^{T_{\text{on}}-1} + (H_{A_2} - H_{A_4}) \left(1 - [\lambda(0)]^{T_{\text{on}}-1}\right)\right] \\ & + \frac{[\lambda(0)]^{NT_\beta}}{1 - [\lambda(0)]^{T_\gamma}} \left[H_{G_1} + H_{G_2} [\lambda(0)]^{T_{\text{on}}-1} + (H_{A_2} - H_{A_3}) \left(1 - [\lambda(0)]^{T_{\text{on}}-1}\right)\right]. \end{aligned} \tag{118}$$

Then, the total energy savings τ become

$$\tau = \tau_\beta + \tau_\gamma. \tag{119}$$

4.2.4.2 Delay

To calculate the expected value $E\{\Gamma\}$ of the packet delay, also called waiting time in queuing theory terminology, we need to compute (a) the expected value of the delay W_1, W_2, W_3,

and W_4 experienced by the packets sent in A_1, A_2, A_3, and A_4, respectively, and (b) the probability of a packet being sent in each of those states. An important fact to keep in mind is that a packet sent through A_i did not necessarily arrive in the previous "sleep" state; instead, it may have arrived while the packets that arrived during the "sleep" state are being sent during A_i. Therefore, a packet sent through A_i may experience an additional amount of delay even if it did not arrive during a "sleep" period.

To compute the delay, we utilize the results from queuing theory establishing the expected value $E\{W\}$ of the packet waiting time in a system with vacation [82]. In such context,

$$E\{W\} = \frac{\lambda^2 E\{X^2\} + bE\{\Lambda^2\} - \rho(\lambda + 1)}{2\lambda(1-\rho)} + \frac{E\{v(v-1)\}}{2E\{v\}}, \tag{120}$$

where v is the length of the vacation, the first term represents the waiting time in a system without vacation, and the second term represents the residual life of the vacation time. In the context of DRX, v corresponds to the amount of time during which the BS buffers packets before entering a continuous reception state. Therefore, v is a deterministic value equal to 1 for A_1 and A_2, to $T_\gamma - T_{\text{on}} + 1$ for A_3, and to $T_\beta - T_{\text{on}} + 1$ for A_4. It then follows that

$$E\{W_1\} = E\{W_2\} = \frac{\lambda^2 E\{X^2\} + bE\{\Lambda^2\} - \rho(\lambda + 1)}{2\lambda(1-\rho)}, \tag{121a}$$

$$E\{W_3\} = E\{W_1\} + \frac{T_\gamma - T_{\text{on}}}{2}, \tag{121b}$$

$$E\{W_4\} = E\{W_1\} + \frac{T_\beta - T_{\text{on}}}{2}. \tag{121c}$$

We now proceed to compute the probability of a packet being sent from state A_i. First, we denote by $\hat{R}_{A_i}$ the number of packets sent during A_i. By Little's theorem,

$$E\{\hat{R}_{A_i}\} = \frac{H_{A_i}}{b} = \frac{1}{1-\rho} E\{R_{A_i}\}. \tag{122}$$

Then, the probability of a packet being sent from A_i is

$$\text{Prob}\{Y = A_i\} = \frac{\pi_{A_i} E\{\hat{R}_{A_i}\}}{\sum_{j=1}^{4} \pi_{A_j} E\{\hat{R}_{A_j}\}} = \frac{\pi_{A_i} H_{A_i}}{\sum_{j=1}^{4} \pi_{A_j} H_{A_j}} = \frac{\pi_{A_i} E\{R_{A_i}\}}{\sum_{j=1}^{4} \pi_{A_j} E\{R_{A_j}\}}, \tag{123}$$

where Y denotes the state from which the packet is sent.

Now, applying the law of total expectation, we have that

$$\begin{aligned} E\{\Gamma\} &= E\{E\{\Gamma|Y\}\} = \sum_{i=1}^{4} E\{\Gamma|Y = A_i\}\text{Prob}\{Y = A_i\} \\ &= \sum_{i=1}^{4} E\{W_i\}\text{Prob}\{Y = A_i\}. \end{aligned} \tag{124}$$

After further simplification, the above expression becomes

$$E\{\Gamma\} = E\{W_1\} + \frac{(T_\gamma - T_{\text{on}})\pi_{A_3} E\{R_{A_3}\} + (T_\beta - T_{\text{on}})\pi_{A_4} E\{R_{A_4}\}}{2\sum_{j=1}^{4} \pi_{A_j} E\{R_{A_j}\}}. \tag{125}$$

As mentioned previously, the first term represents the waiting time in a system with no vacation/DRX. Hence, the second term denotes the additional waiting time caused by the use of DRX.

4.2.5 Performance Evaluation

As shown in Table 5, the operation of DRX depends on multiple parameters, in addition to the arrival and departure PMFs, creating a large number of possible parameter combinations. Therefore, the focus of this section is to show the validity of our modeling approach, key insights about the role of each DRX parameter, the achievable energy savings, and the impact on the packet delay.

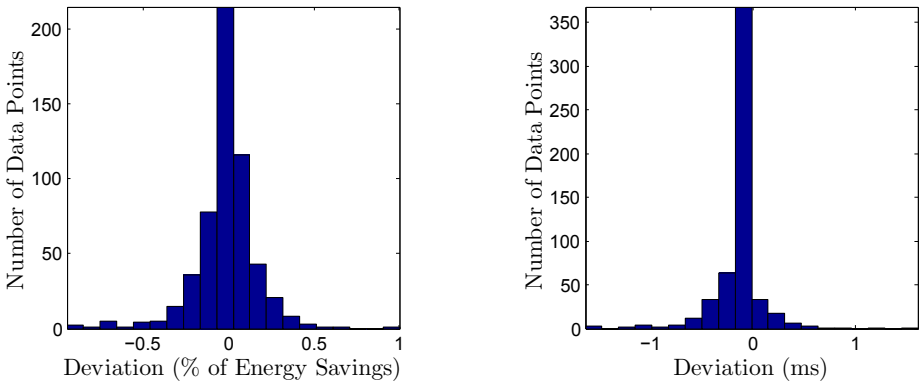
To validate our model and the performance of DRX, we simulated a DRX system consisting of a BS and a UE. The UE supports all the DRX states and parameters described in Section 4.2.1. The BS has an infinite buffer, as typically considered in the literature,

where it stores any packet that cannot be immediately sent to the UE because the UE is in a "sleep" state or the previously buffered packets are being sent. Once the UE is "awake," the buffered packets are sent by the BS following a First In, First Out (FIFO) scheme. We consider that there is no packet loss or retransmissions between the BS and the UE. We consider that the number of packets Λ that arrive in a single subframe follows a Poisson distribution with parameter λ. For the service time X, we utilize a modified Poisson distribution:

$$\text{Prob}\{X = k|b\} = (b-1)^{k-1}\frac{e^{-(b-1)}}{(k-1)!}, \qquad b > 1; k = 1, 2, 3, ... \tag{126}$$

Therefore, $E\{X\} = b$. X can be interpreted as adding 1 to the result of generating a random variable from a Poisson distribution whose mean is $b - 1$.

In Figure 26, we depict the histogram of the deviation between the analytical and simulation results for the energy savings and delay metrics across multiple DRX parameters combinations. For each combination, the operation of the LTE DRX during 1 million subframes was simulated. At the end of each simulation, the energy savings and delay metrics were computed and compared to the results found from the analytical expressions. For each configuration, the deviation is then computed as the difference between the theoretical and simulated performance metrics.



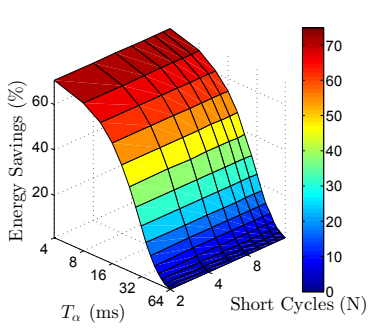
(a) Energy savings.

(b) Delay.

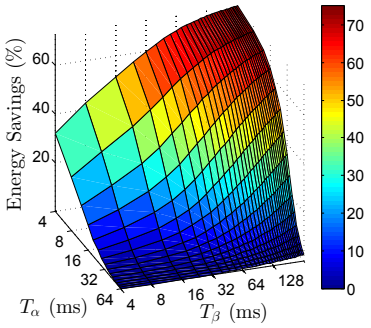
Figure 26: Deviation of theoretical from experimental metrics for the LTE DRX with $\lambda = 0.1$, $b = 2.5$ms, $T_\alpha \in [4, 8, 16, 32, 64]$ms, $N \in [2, 4, 8, 16]$, $T_\beta \in [4, 8, 16, 32, 64, 128, 256]$ms, $T_{\text{on}} \in [2, 4, 8, 16, 32, 64, 128]$ms, and $T_\gamma = 2T_\beta$.

From Figure 26, we observe that the deviation for both metrics is extremely low. In particular, the absolute deviation in the energy savings is less than 1% of energy savings. Similarly, the absolute deviation in the delay is mostly within 1ms. From these results, we validate the significantly high accuracy of the analytical expressions derived for the performance metrics. This validation allows us to further examine the performance metrics directly through their analytical expressions.

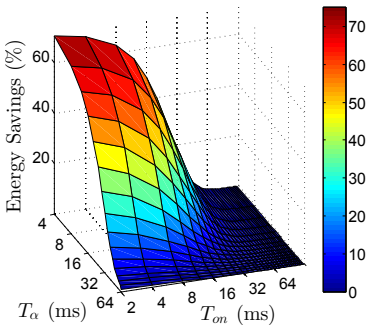
In Figure 27, we depict the analytical results for the energy savings. From Figure 27a, 27b, and 27c, we observe that the length of the inactivity timer has a significant impact on the energy savings. Particularly, smaller values of such timer increase the energy savings. This behavior occurs because a shorter inactivity period increases the probability of entering the DRX cycles. From Figure 27b, 27d, and 27e, we observe that the energy savings increase when the length of the short DRX increases. However, such improvements in energy savings are hindered when the length of the "on" period is increased because that period decreases the amount of time a UE spends in a "sleep" period.



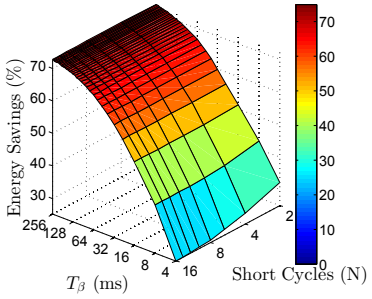
(a) Inactivity timer vs. short cycles. $T_\beta = 128$ms, $T_{\text{on}} = 2$ms.



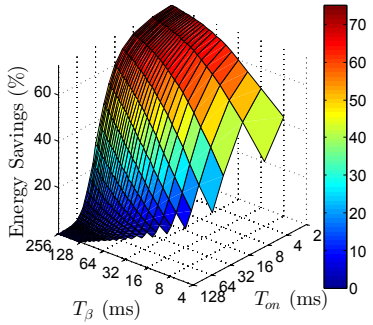
(b) Inactivity timer vs. short DRX length. $T_{\text{on}} = 2$ms, $N = 2$.



(c) Inactivity timer vs. "on" period length. $T_\beta = 128$ms, $N = 2$.



(d) Short DRX Length vs. short cycles. $T_\alpha = 4$ms, $T_{\text{on}} = 2$ms.

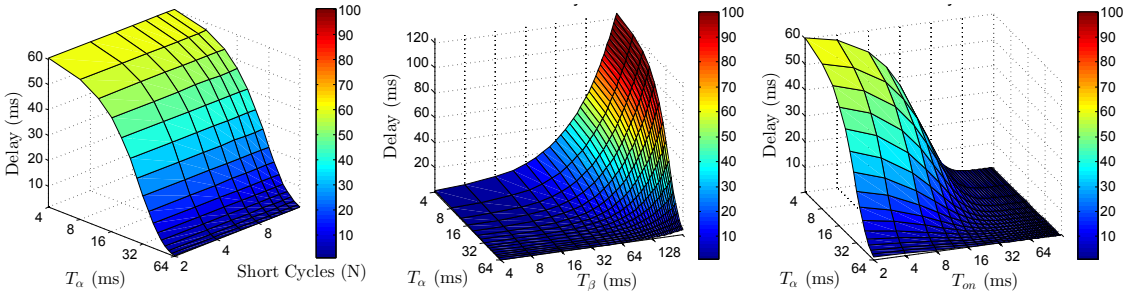


(e) Short DRX length vs. "on" period length. $T_\alpha = 4$ms, $N = 2$.

Figure 27: Energy savings in LTE DRX with parameters $\lambda = 0.1, b = 2.5$ms, and $T_\gamma = 2T_\beta$.

The impact of the DRX parameters on the packet delay is shown in Figure 28. Each graph in Figure 28 was generated using the same set of parameters as those in Figure 27. As expected, the highest level of energy savings corresponds to the greatest packet delay. Nevertheless, by comparing Figure 27b and Figure 28b, we observe that there is a subset of values that can be assigned to T_β and T_α that provides significant energy savings without severely increasing the packet delay. For example, $T_\beta = 32$ms and $T_\alpha \geq 16$ms provide energy savings between 30% and 60% while keeping the packet delay below 20ms. From Figure 27e and Figure 28e, we observe that a similar region exists for T_β and T_{on}. For example, $T_\beta = 32$ms and $T_{\text{on}} \leq 4$ms provide energy savings between 50% and 60% while the packet delay remains below 15ms. In general, we also notice that the number of short DRX cycles had a smaller impact on the energy savings and packet delay, as compared to

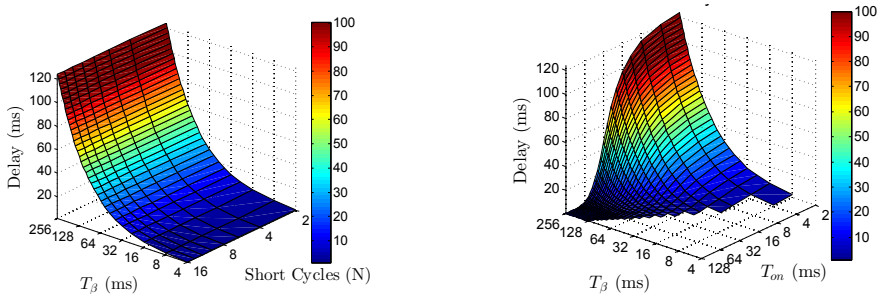
the rest of the DRX parameters.



(a) Inactivity timer vs. short cycles. $T_\beta = 128$ms, $T_{\text{on}} = 2$ms.

(b) Inactivity timer vs. short DRX length. $T_{\text{on}} = 2$ms, $N = 2$.

(c) Inactivity timer vs. "on" period length. $T_\beta = 128$ms, $N = 2$.

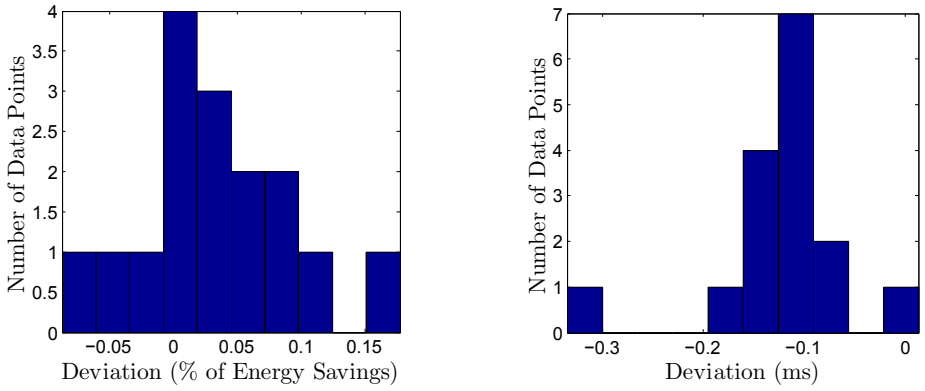


(d) Short DRX length vs. short cycles. $T_\alpha = 4$ms, $T_{\text{on}} = 2$ms.

(e) Short DRX length vs. "on" period length. $T_\alpha = 4$ms, $N = 2$.

Figure 28: Waiting time, i.e., delay, in LTE DRX with parameters $\lambda = 0.1, b = 2.5$ms, and $T_\gamma = 2T_\beta$.

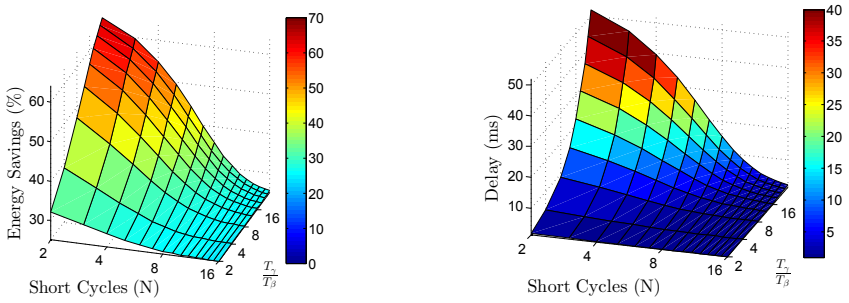
In Figure 29, we depict the histogram of the deviation between the analytical and simulation results for the energy savings and delay metrics as N and $\frac{T_\gamma}{T_\beta}$ are varied. For each combination of the DRX parameters, the deviation was computed the same way as for Figure 26. Here, we also have extremely low deviations. In particular, the absolute deviation in the energy savings is less than 0.2% of energy savings. Similarly, the absolute deviation in the delay is mostly within 0.2ms. These low levels of deviation allow us to further examine the performance metrics, as N and $\frac{T_\gamma}{T_\beta}$ are varied, directly through the analytical expressions.



(a) Energy savings. (b) Delay.

Figure 29: Deviation of theoretical from experimental metrics for LTE DRX with $\lambda = 0.1$, $b = 2.5$ms, $T_\alpha = 4$ms, $T_\beta = 4$ms, $T_{\text{on}} = 2$ms, $N \in [2, 4, 8, 16]$, and $\frac{T_\gamma}{T_\beta} \in [2, 4, 8, 16, 32]$.

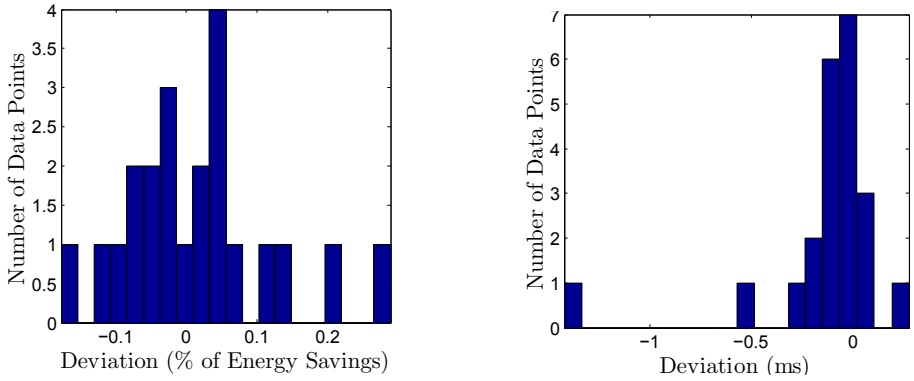
In Figure 30, we present the energy savings and delay metrics for the case when the number of short DRX cyles (N) and the ratio $\frac{T_\gamma}{T_\beta}$ are adjusted. We observe that a decreasing N augments the energy savings. Such relationship is formed because a lower number of short DRX cycles gives the UE a higher probability of reaching the long DRX cycle and, therefore, increases the amount of time spent in the "sleep" period. We also notice that the impact of reducing N increases as the ratio $\frac{T_\gamma}{T_\beta}$ augments. This result follows from the fact that having a significantly larger "sleep" period in the long DRX compared to the one of the short DRX increases the ratio of energy saved in a single long DRX cycle to the energy saved in a single short DRX cycle.



(a) Energy savings. (b) Waiting time, i.e., delay.

Figure 30: Short cycles (N) vs. $\frac{T_\gamma}{T_\beta}$ in LTE DRX with parameters $\lambda = 0.1, b = 2.5$ms, $T_\alpha = 4$ms, $T_\beta = 4$ms, and $T_{\text{on}} = 2$ms.

In Figure 31, we depict the histogram of the deviation between the analytical and simulation results for the energy savings and delay metrics as λ and b are varied. For each combination of the DRX parameters, the deviation was computed as similarly done for Figure 26. Here, we also have extremely low deviations. In particular, the absolute deviation in the energy savings is less than 0.3% of energy savings. Similarly, the absolute deviation in the delay is mostly within 0.5ms. These low levels of deviation allow us to further examine the performance metrics, as λ and b are varied, directly through the analytical expressions.

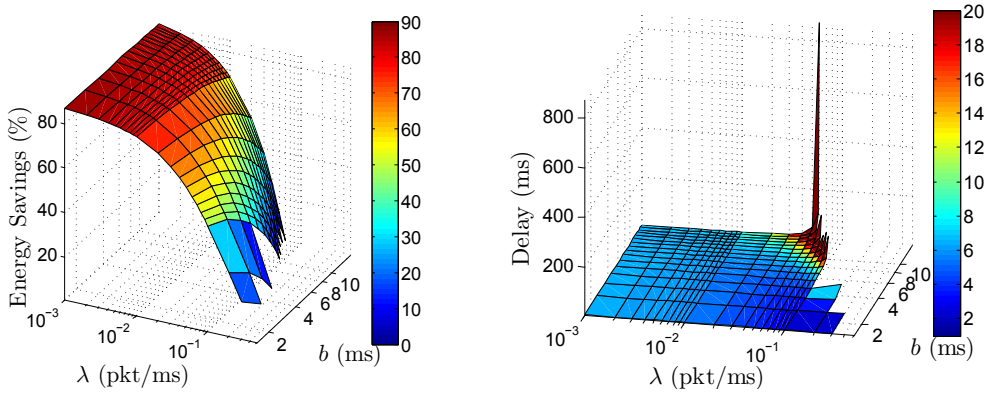


(a) Energy savings. (b) Delay.

Figure 31: Deviation of theoretical from experimental metrics for the LTE DRX with $\lambda \in [0.1, 0.05, 0.01, 0.001]$, $b \in [1.5, 2.5, 4.5, 6.5, 8.5, 16.5]$ms, $T_\alpha = 4$ms, $T_\beta = 8$ms, $T_{\text{on}} = 2$ms, $N = 2$, and $T_\gamma = 2T_\beta$.

In Figure 32, we present the energy savings and delay metrics for the case when the mean value for the number of packets that arrive in a single subframe and the mean service time are adjusted. As b and λ increase, the energy savings decrease. However, the waiting time follows a different pattern. For small values of b, e.g., 1.5ms and 2.5ms, and increasing λ, the waiting time decreases. This behavior occurs because more packets arrive during a continuous reception state and, therefore, are not affected by the additional delay caused by the "sleep" period. It is also true that during the "sleep" period more packets will arrive and be buffered, which will affect not only their service time, but also that of future packets. However, the second effect is mitigated by a low value of b. On other hand, for large

values of b, such mitigation is not possible, and the overall packet waiting time increases significantly. This phenomenon is seen in Figure 30b as $\lambda \rightarrow 0.05, b \rightarrow 16.5$ms and $\lambda \rightarrow 0.1, b \rightarrow 8.5$ms, i.e., $\rho \rightarrow 1$. From the previous observations, we can state that for a given set of DRX parameters, the impact on the energy savings and the packet delay can be drastically affected not only by the mean packet arrival rate, but also by the mean service time, particularly when the overall load ρ is large. This phenomenon can reach the point where both the packet delay and the energy savings deteriorate, contrary to the commonly assumed behavior where the improvement of one causes a worsening of the other, as shown in Figure 27 and Figure 28.



(a) Energy savings.

(b) Waiting time, i.e., delay.

Figure 32: Mean packet arrival/subframe vs. mean service time in LTE DRX with parameters $T_\alpha = 4$ms, $T_\beta = 8$ms, $T_\gamma = 16$ms, $T_{\text{on}} = 2$ms, and $N = 2$.

4.3 Cross-Carrier-Aware DRX Analysis

With the introduction of single- and multi-stream carrier aggregation, an LTE-A UE connects simultaneously to up to five CCs, thus increasing its energy consumption up to five times. For such UE, DRX still remains the most viable option to reduce the energy consumption. Nevertheless, the existing literature has focused on using either the same DRX parameters for all CCs or completely different parameters for each CC [70][75][81]. The first approach is simple, but extremely rigid and inefficient, since all the CCs are active, regardless of whether any traffic is transmitted. On the other hand, the second approach

provides flexibility per CC, but completely discards the cross-carrier awareness that exists at the BS and the UE. However, by exploiting such awareness, it is possible to further increase the benefits of DRX, as we will describe.

The cross-carrier awareness arises from the fact that the medium access control (MAC) [70] is exposed to the multi-CC nature of the physical layer (PHY), even though the radio link control (RLC) [83] and the packet data convergence protocol (PDCP) [84] are unaware of such nature, as shown in Figure 33 for the downlink at the BS [85].

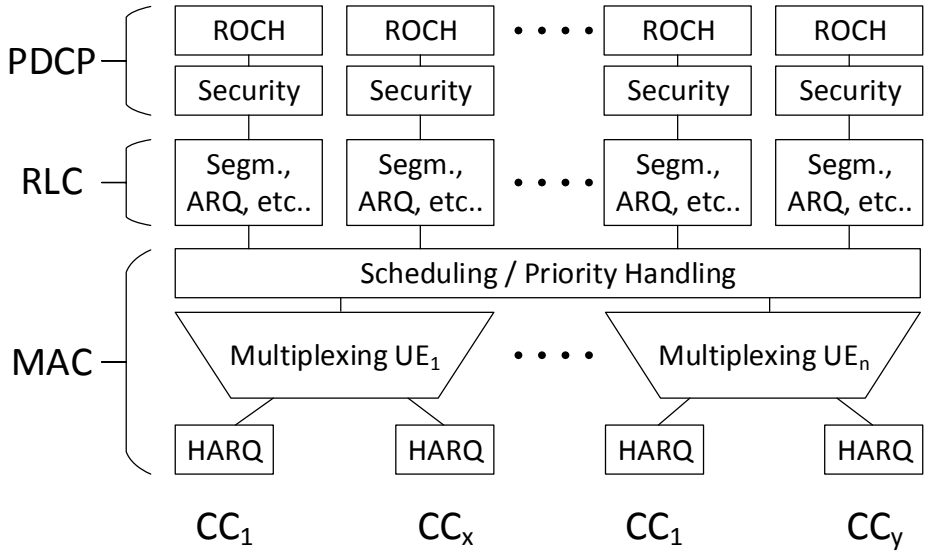


Figure 33: Impact of carrier aggregation on the base station downlink protocol stack.

In a scenario supporting carrier aggregation, the CC that is used for establishing the radio resource control (RRC) connection [86] is called the primary cell (PCell). It provides non-access-stratum (NAS) [87] mobility information, as well as security input. If a UE supports multiple CCs, then additional secondary cells (SCells) can be added to the user's set of serving cells. One important aspect is that the PCell and SCell of different UEs need not be the same, i.e., the SCell of one UE may be the PCell of another UE.

To exploit the cross-carrier awareness, we take advantage of the scheduling methods supported at the MAC layer. There, both same- and cross-carrier scheduling methods are

supported. The same-carrier method corresponds to the scheme used for LTE. The cross-carrier method allows the BS to use a particular CC to assign resources contained in a different CC. However, there are certain restrictions. First, the PCell is the only one that schedules its resources. It does so through the Physical Downlink Control Channel (PDCCH), at the PHY. Second, cross-carrier scheduling only applies when the PDCCH of a SCell is not configured. As a result of these restrictions, a CC whose resources are allocated through cross-carrier scheduling cannot have its own DRX parameters. Conversely, a CC that has its own PDCCH and, therefore, its own DRX parameters, cannot be scheduled by the PCell through cross-carrier scheduling. We will now describe our cross-carrier-aware DRX that accounts for the aforementioned constraints and provides improved performance compared to the traditional DRX.

The basic concepts behind our cross-carrier-aware DRX are captured in Figure 34. This figure depicts the operation of a UE that has three CCs, where CC1 represents the PCell acting as an anchor CC and the remaining CCs acting as SCells.

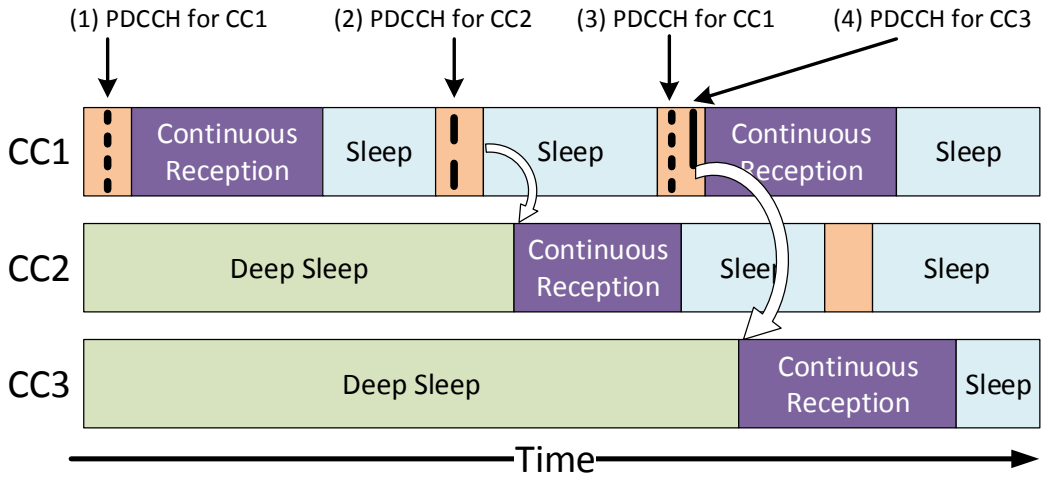


Figure 34: Cross-carrier-aware DRX operation.

1. Event 1: At the UE, the anchor CC receives a PDCCH indicating that the upcoming subframes in CC1 contain data. Following the traditional DRX behavior, CC1 enters a continuous reception state until no more data is received and then returns to “sleep”. While CC1 performs the aforementioned steps, CC2 and CC3 remain in a “deep

sleep" state since no data has arrived for those CCs. In contrast to the "sleep" states in the traditional LTE DRX, the "deep sleep" in our proposed solution does not require CC2 and CC3 to "wake up" to check if a PDCCH has arrived. Therefore, our solution has a greater potential to save energy.

2. Event 2: Similar to the concept of cross-carrier scheduling, we propose the use of a cross-carrier DRX activation. For example, once the UE "wakes up" its anchor CC to listen for the presence of a PDCCH, such UE identifies whether a PDCCH indicates subsequent data in the anchor CC or in a SCell. In event 2, the PDCCH indicates subsequent data in CC2. Therefore, the UE "wakes up" CC2 from the "deep sleep" state so that it can receive the upcoming data, while allowing CC1 to go back to its normal "sleep" state. Because of the synchronization and timing differences between CC1 and CC2, the activation of the latter is delayed, and so is the packet reception. This delay is the penalty that our system incurs to provide the energy savings of the "deep sleep" state. In event 2, we also observe that the proposed cross-carrier activation is selective, e.g., in event 2, only CC2 is activated, leaving CC3 in the "deep sleep" state. Another feature of our proposed solution is the ability to set per-carrier DRX parameters. For example, once CC2 is activated as a result of event 2, we allow for CC2 to utilize its own DRX parameters, which may be completely different from the ones of CC1, and, therefore, can be optimized to the characteristics of the traffic carried over CC2.

3. Events 3 and 4: At the UE, the anchor CC simultaneously receives two PDCCH. The first PDCCH indicates that the upcoming subframes in CC1 contain data; therefore, CC1 then enters a continuous reception state. The second PDCCH indicates that the subsequent subframes in CC3 contain data; therefore, CC3 then enters a continuous reception state. We observe that the operation of CC2 is not affected by the DRX events of CC1 since at this point CC2 is following its own DRX parameters. As

in event 2, once CC3 enters the continuous reception state, it follows its own DRX parameters, which can be different from the ones of CC1 and CC2.

4.3.1 Cross-Carrier-Aware DRX Model

As previously described, the anchor CC in our proposed cross-carrier-aware DRX follows the traditional LTE DRX operation and supports the cross-carrier-aware DRX operation of the SCells. Such triggering does not affect the DRX operation of the anchor; therefore, its analysis and performance metrics correspond to the ones presented in Section 4.2. In this section, we focus on the analysis of the cross-carrier-aware DRX operation of a single SCell. Such operation is captured as a finite state machine (FSM) model in Figure 35. Compared to the FSM of the traditional LTE DRX, depicted in Figure 24, the one in Figure 35 does not have a long DRX cycle that may be repeated any number of times; instead, it has a single entrance to the "deep sleep" state and a subsequent exit only to the continuous reception state. The exit transition from the "deep sleep" state is triggered by the events occurring at the anchor CC.

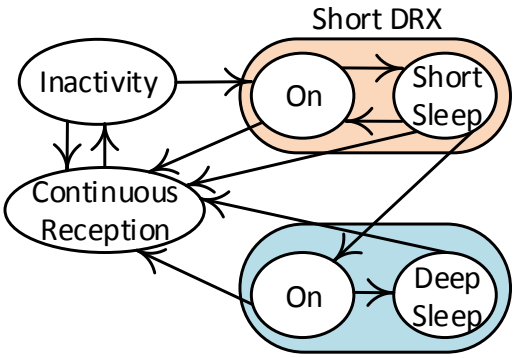


Figure 35: Cross-carrier-aware DRX finite state machine for SCell.

As discussed in Section 4.2.1, the DRX operation cannot be directly modeled as a typical Markov Chain using the FSM of Figure 35 because

- the amount of time spent in each state differs among states;
- the short DRX cycle is repeated up to M times before the UE exits it. Therefore,

memory is required to keep track of the number of cycles that have been previously executed;

- the amount of time spent in the continuous reception state depends on the previously executed state.

We utilize a semi-Markov Chain with late arrival to model the cross-carrier-aware DRX operation of the SCell, as shown in Figure 36. The description of each state is shown in Table 6. The inactivity period is modeled by the state R. Each of the M possible short DRX cycles is explicitly modeled as a pair of Y_{2i} and Y_{2i-1} states, $i \in [1, M]$. The former represents the "on" period of the i-th short DRX cycle, while the latter represents the "sleep" period of the same cycle. State V_1 represents an "on" period executed after M short DRX cycles and before entering the "deep sleep" state V_2. Even though the "deep sleep" state is represented in Figure 36 as a single state V_2, it encompasses a group of states that captures the events occurring at the anchor while the SCell is in the "deep sleep".

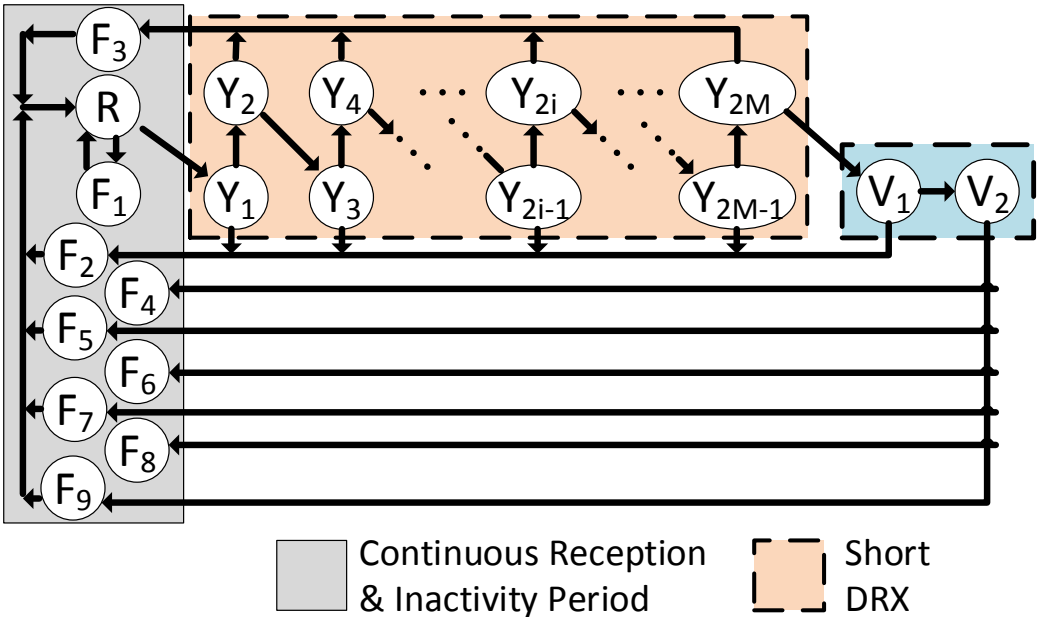


Figure 36: Cross-carrier-aware DRX semi-Markov Chain model for SCell.

Table 6: Cross-carrier-aware DRX states description.

State	Description
R	Inactivity period.
Y_{2i} $i \in [1, M]$	"Sleep" period of the i-th short DRX cycle.
Y_{2i-1} $i \in [1, M]$	"On" period of the i-th short DRX cycle.
V_1	"On" period preceding the "deep sleep" state.
V_2	"Deep sleep" state.
F_1	Continuous reception following state R.
F_2	Continuous reception following states Y_{2i-1} and V_1.
F_3	Continuous reception following state Y_{2i}.
F_i $i \in [4, 9]$	Continuous reception following state V_2.

Rather than using a single state to model the continuous reception, we utilize nine different states. The reasoning behind this approach is that the number of packets sent by the BS during continuous reception and, therefore, the amount of time spent in continuous reception depend on the SCell state at the moment that the packet triggering the continuous reception arrives at the BS. For example, the expected number of packets received at the BS during the "sleep" period of a short DRX cycle is larger than that received during the "on" period of a short DRX cycle. Therefore, the expected duration of state F_3, defined in Table 4, is longer than that of F_2. From Figure 36, we also observe that state V_2 leads to each of the five continuous reception states $F_i, i \in [4, 9]$. This occurs because V_2 encompasses multiple internal states, and each of them may produce a different number of packets that need to be sent in a continuous reception state.

Except for the transitions that exit the "deep sleep" state, all the transitions in Figure 36 are controlled by the parameters in Table 7 and reflect the DRX operation described in Section 4.3. Therefore, the same events that trigger a transition in the traditional LTE DRX (Section 4.2.1) determine the transitions (i) from the continuous reception states to the

inactivity state and vice versa, (ii) within the short DRX cycles, (iii) from the short DRX cycles to the continuous reception states, and (iv) from Y_{2M} to the next "on" period. The transitions that exit the "deep sleep" state V_2 are controlled by the events occurring at the anchor CC and its DRX parameters, which are shown in Table 8[14].

Table 7: Cross-carrier-aware DRX parameters for the SCell.

Parameter	Description
$T_{\alpha 2}$	Inactivity period length.
$T_{\beta 2}$	Short DRX cycle length.
M	Number of short DRX cycles.
$T_{\text{on}2}$	"On" period length.

Table 8: Cross-carrier-aware DRX parameters for the anchor CC.

Parameter	Description
T_{α_1}	Inactivity period length.
T_{β_1}	Short DRX cycle length.
T_{γ_1}	Long DRX cycle length.
N	Number of short DRX cycles.
$T_{\text{on}1}$	"On" period length.

Once the SCell enters the "deep sleep" state, the UE starts tracking the events of the anchor CC waiting for a PDCCH indicating that the BS has a packet ready for the SCell. In other words, entering the "deep sleep" state is equivalent to starting the semi-Markov Chain of the anchor CC from a randomly selected state, as depicted in Figure 37.

[14]The parameters in Table 8 are equivalent to the ones in Table 5. Here, the notation has been adjusted to facilitate the differentiation of the parameters of the anchor CC from the ones of the SCell.

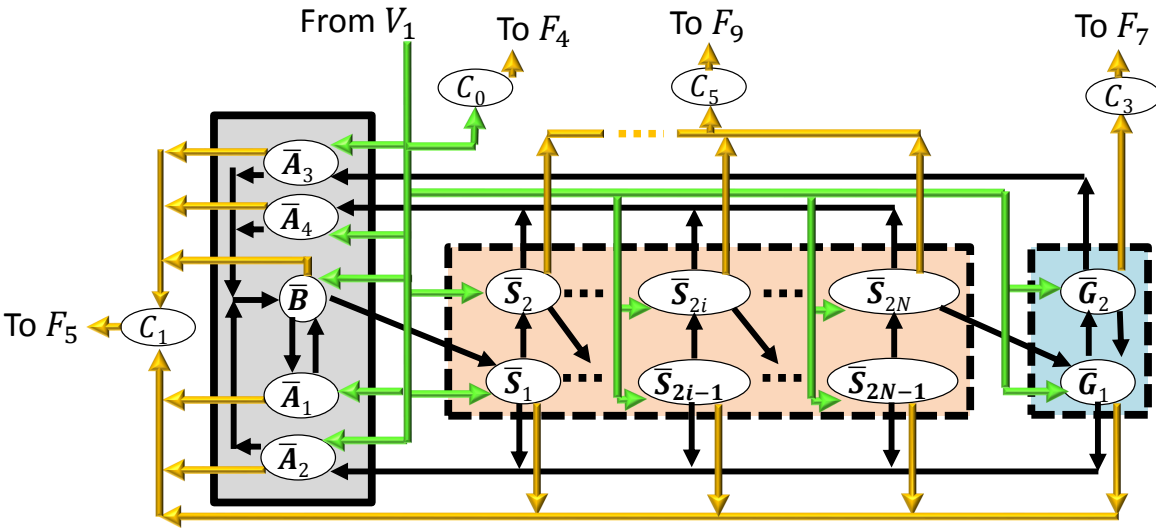


Figure 37: Internal semi-Markov Chain of the "deep sleep" state.

In Figure 37, the states of the form $\bar{U}$ represent the ones being tracked at the anchor CC. States $C_i, i \in \{0, 1, 3, 5\}$, capture the ones executed after the anchor CC receives a PDCCH indicating that the BS has a packet for the SCell. After any such state is executed, the "deep sleep" state ends, and the SCell transitions to a continuous reception state F_i.

When the SCell transitions from V_1 to V_2, not only the anchor CC is executing a randomly selected state, but also the amount of time left until the end of such state is a random variable. For example, when the SCell enters the "deep sleep" state, it may occur that the anchor CC is in the middle of a "sleep" period of a short DRX rather than at the beginning of such period. For this reason, when the SCell enters the "deep sleep" state, it needs to synchronize with the anchor CC to account for the randomness in the amount of time left until the end of the landed state. Such synchronization is captured in Figure 38 by introducing the states of the form $\hat{U}$. States C_2 and C_4 are also added to account for the possibility of exiting the "deep sleep" state during or immediately after a synchronization state associated with a "sleep" state.

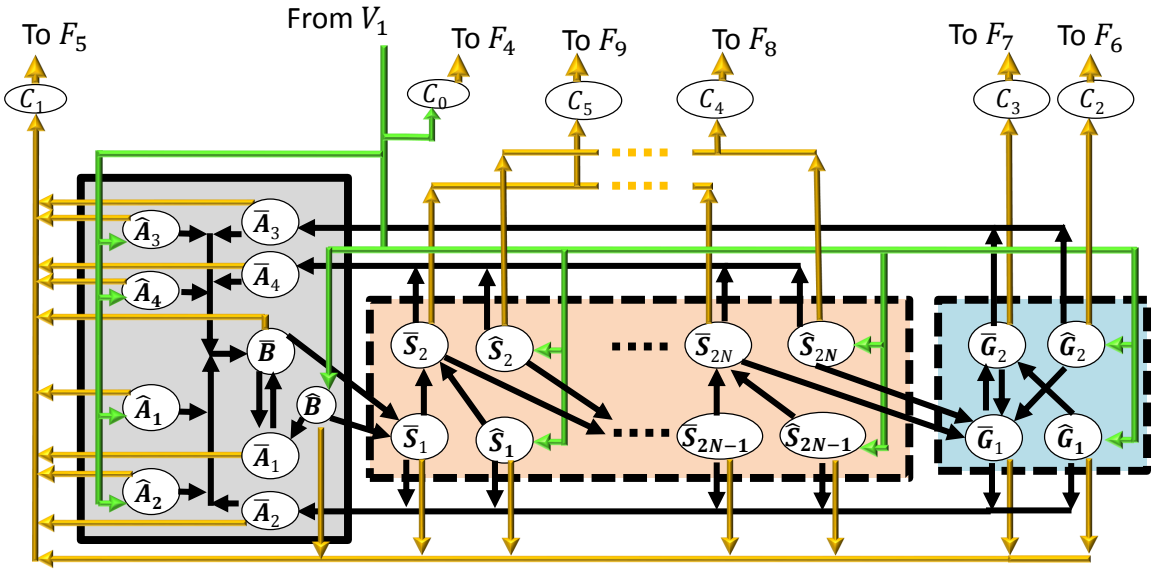


Figure 38: Internal semi-Markov Chain of the "deep sleep" state with synchronization states.

In Figure 38, C_0 has a key role in the transition probabilities from V_1 to the synchronization states. C_0 captures the case when a packet arrives at the BS for the SCell during the last subframe x of V_1, i.e., when the transition to the "deep sleep" state is inevitable. Since that packet cannot be scheduled to be sent during subframe x, it must wait at least for the next subframe $x + 1$. If during subframe $x + 1$ the anchor is in a non-"sleep" state, then the BS can notify the UE in that same subframe that it has a packet ready for the SCell; therefore, that the SCell should exit the "deep sleep" state. This scenario is the one captured by state C_0. As a consequence of this setup, the only cases where a synchronization state is reached are

- when no packet arrived to the BS for the SCell during the last subframe x of V_1, or
- when the anchor CC is in a "sleep" state during subframe $x + 1$.

4.3.2 Stationary Probability - Embedded Markov Chain - SCell

In this section, we present the calculation of the stationary probability of the EMC corresponding to the semi-Markov Chain in Figure 36.

Utilizing the notation $p_{U,U'}$ for the transition probability from state U to state U' and π_U for the stationary probability of state U in the EMC, the stationary probabilities π of the EMC follow these relationships:

$$\pi_R = \sum_{i=1}^{9} \pi_{F_i}, \tag{127a}$$

$$\pi_{Y_1} = \pi_R p_{R,Y_1}, \tag{127b}$$

$$\pi_{Y_i} = \pi_{Y_{i-1}} p_{Y_{i-1},Y_i}, \qquad i \in [2, 2M], \tag{127c}$$

$$\pi_{V_1} = \pi_{Y_{2M}} p_{Y_{2M},V_1}, \tag{127d}$$

$$\pi_{V_2} \triangleq \sum_{\forall U \in V_2} \pi_U, \tag{127e}$$

$$\pi_{F_1} = \pi_R p_{R,F_1}, \tag{127f}$$

$$\pi_{F_2} = \sum_{i=1}^{M} \pi_{Y_{2i-1}} p_{Y_{2i-1},F_2} + \pi_{V_1} p_{V_1,F_2}, \tag{127g}$$

$$\pi_{F_3} = \sum_{i=1}^{M} \pi_{Y_{2i}} p_{Y_{2i},F_3}. \tag{127h}$$

From these expressions we obtain:

$$\pi_R = \frac{\pi_{Y_1}}{p_{R,Y_1}}, \tag{128a}$$

$$\pi_{Y_i} = \pi_{Y_1} \prod_{j=2}^{i} p_{Y_{j-1},Y_j}, \qquad i \in [2, 2M], \tag{128b}$$

$$\pi_{V_1} = \pi_{Y_1} \omega_2, \tag{128c}$$

$$\pi_{V_2} = \pi_{Y_1} \omega_2 \omega_3, \tag{128d}$$

$$\pi_{F_1} = \pi_{Y_1} \frac{p_{R,F_1}}{p_{R,Y_1}}, \tag{128e}$$

$$\pi_{F_2} = \pi_{Y_1} \left[p_{Y_1,F_2} + \omega_2 p_{V_1,F_2} + \sum_{i=2}^{M} \left(p_{Y_{2i-1},F_2} \prod_{j=2}^{2i-1} p_{Y_{j-1},Y_j} \right) \right], \tag{128f}$$

$$\pi_{F_3} = \pi_{Y_1} \sum_{i=1}^{M} \left(p_{Y_{2i},F_3} \prod_{j=2}^{2i} p_{Y_{j-1},Y_j} \right), \tag{128g}$$

where

$$\omega_2 = p_{Y_{2M},V_1} \prod_{j=2}^{2M} p_{Y_{j-1},Y_j}, \qquad \omega_3 = \sum_{\forall U \in V_2} \frac{\pi_U}{\pi_{V_1}}. \tag{129}$$

With these expressions, the value of π_{Y_1} becomes

$$\pi_{Y_1} = \left[1 + \frac{2}{p_{R,Y_1}} + \omega_2(1+\omega_3) + \sum_{i=2}^{2M} \prod_{j=2}^{i} p_{Y_{j-1},Y_j} \right]^{-1}, \tag{130}$$

which can be plugged into Eq. (128) to obtain the stationary probabilities of all other states of the EMC model.

As in Section 4.2.2, we utilize a late-arrival model for the packet arrival in the SCell. We utilize Λ_2 to denote the number of packets that arrive in a single subframe. The PMF of Λ_2 is defined as

$$\lambda_2(k) \triangleq \text{Prob}\{\Lambda_2 = k\}, \qquad k = 0, 1, 2..., \tag{131}$$

and its mean value is defined as

$$\lambda_2 \triangleq E\{\Lambda_2\}, \tag{132}$$

where $E\{\Lambda_2\}$ denotes the expected value of Λ_2. We denote by X_2 the service time (measured in subframes) of a single packet in the SCell. The PMF of X_2 is defined as

$$b_2(l) = \text{Prob}\{X_2 = l\}, \qquad l = 1, 2, ..., \tag{133}$$

and its mean value is defined as

$$b_2 \triangleq E\{X_2\}. \tag{134}$$

Having the arrival model, the service discipline, and the PMF of the packet arrivals and the service time, we proceed to characterize the transition probabilities. To achieve this, we (a) apply the conditions that trigger each transition, as described in Section 4.3.1, and (b) consider that the probability of a BS receiving no packets in a time period of length v is equal to $[\lambda_2(0)]^v$ since the number of packet arrivals in successive subframes constitutes

a sequence of i.i.d. random variables. Consequently, the probability that the BS receives at least one packet in a time period of length v is equal to $1 - [\lambda_2(0)]^v$.

$$p_{R,Y_1} = [\lambda_2(0)]^{T_{\alpha 2}}, \tag{135a}$$

$$p_{R,F_1} = 1 - [\lambda_2(0)]^{T_{\alpha 2}}, \tag{135b}$$

$$p_{Y_{2i-1},Y_{2i}} = [\lambda_2(0)]^{T_{\text{on}2}-1}, \qquad i \in [1, M], \tag{135c}$$

$$p_{Y_{2i-1},F_2} = 1 - [\lambda_2(0)]^{T_{\text{on}2}-1}, \qquad i \in [1, M], \tag{135d}$$

$$p_{Y_{2i},Y_{2i+1}} = [\lambda_2(0)]^{T_{\beta 2}-T_{\text{on}2}+1}, \qquad i \in [1, M-1], \tag{135e}$$

$$p_{Y_{2i},F_3} = 1 - [\lambda_2(0)]^{T_{\beta 2}-T_{\text{on}2}+1}, \qquad i \in [1, M], \tag{135f}$$

$$p_{Y_{2M},V_1} = [\lambda_2(0)]^{T_{\beta 2}-T_{\text{on}2}+1}, \tag{135g}$$

$$p_{V_1,F_2} = 1 - [\lambda_2(0)]^{T_{\text{on}2}-1}, \tag{135h}$$

$$p_{V_1,V_2} \triangleq \sum_{\forall U \in V_2} p_{V_1,U} = [\lambda_2(0)]^{T_{\text{on}2}-1}, \tag{135i}$$

where p_{V_1,V_2} is not a real transition probability since V_2 represents a group of states, but the probability of going from V_1 to any of the states that belong to V_2. Having the above transition probabilities, we plug them in Eq. (128) and Eq. (129) and obtain

$$\omega_2 = [\lambda_2(0)]^{MT_{\beta 2}}, \tag{136}$$

$$\pi_{Y_1} = \left[\frac{2}{[\lambda_2(0)]^{T_{\alpha 2}}} + \left(1 + [\lambda_2(0)]^{T_{\text{on}2}-1}\right)\left(\frac{1 - [\lambda_2(0)]^{MT_{\beta 2}}}{1 - [\lambda_2(0)]^{T_{\beta 2}}}\right) + (1 + \omega_3)\,[\lambda_2(0)]^{MT_{\beta 2}} \right]^{-1}. \tag{137}$$

By denoting $\phi_2 \triangleq \pi_{Y_1}$, the expressions for the stationary probabilities for the rest of the states become

$$\pi_R = \frac{\phi_2}{[\lambda(0)]^{T_{\alpha 2}}}, \tag{138a}$$

$$\pi_{Y_i} = \phi_2 \begin{cases} [\lambda_2(0)]^{T_{\beta 2}(i-1)/2} & : i \text{ is odd} \\ [\lambda_2(0)]^{T_{\beta 2}(i-2)/2}\,[\lambda_2(0)]^{T_{\text{on}2}-1} & : i \text{ is even} \end{cases}, i \in [1, 2M], \tag{138b}$$

$$\pi_{V_1} = \phi_2 \left[\lambda_2(0)\right]^{MT_{\beta 2}}, \tag{138c}$$

$$\pi_{V_2} = \phi_2 \left[\lambda_2(0)\right]^{MT_{\beta 2}} \omega_3, \tag{138d}$$

$$\pi_{F_1} = \phi_2 \left(\frac{1}{\left[\lambda_2(0)\right]^{T_{\alpha 2}}} - 1\right), \tag{138e}$$

$$\pi_{F_2} = \phi_2 \left(1 - \left[\lambda_2(0)\right]^{T_{\text{on}2}-1}\right)\left(1 + \frac{1 - \left[\lambda_2(0)\right]^{MT_{\beta 2}}}{1 - \left[\lambda_2(0)\right]^{T_{\beta 2}}} \left[\lambda_2(0)\right]^{T_{\beta 2}}\right), \tag{138f}$$

$$\pi_{F_3} = \phi_2 \left[1 - \left[\lambda_2(0)\right]^{MT_{\beta 2}}\right]\left[1 - \frac{1 - \left[\lambda_2(0)\right]^{T_{\text{on}2}-1}}{1 - \left[\lambda_2(0)\right]^{T_{\beta 2}}}\right], \tag{138g}$$

$$\pi_{F_i} = \phi_2 \left[\lambda_2(0)\right]^{MT_{\beta 2}} \sum_{\forall U \in V_2} \frac{\pi_U}{\pi_{V_1}} p_{U,F_i}, \qquad i \in [4, 9]. \tag{138h}$$

All the expressions above depend on both the internal states of V_2 (Figure 38) and their transition probabilities to the continuous reception states $F_i, i \in [4, 9]$. As such, we now analyze both of these factors.

4.3.3 Stationary Probability - Embedded Markov Chain - "Deep Sleep" Internal States

In this section, we present the calculation of the stationary probabilities of the EMC corresponding to the semi-Markov Chain in Figure 38.

For the synchronization states, such probabilities follow these relationships:

$$\pi_{\hat{B}} = \pi_{V_1} p_{V_1,\hat{B}}, \tag{139a}$$

$$\pi_{\hat{S}_i} = \pi_{V_1} p_{V_1,\hat{S}_i}, \qquad i \in [1, 2N], \tag{139b}$$

$$\pi_{\hat{G}_i} = \pi_{V_1} p_{V_1,\hat{G}_i}, \qquad i \in [1, 2], \tag{139c}$$

$$\pi_{\hat{A}_i} = \pi_{V_1} p_{V_1,\hat{A}_i}, \qquad i \in [1, 4]. \tag{139d}$$

For the non-synchronization states, such probabilities follow these relationships:

$$\pi_{\bar{S}_1} = \pi_{\bar{B}} p_{\bar{B},\bar{S}_1} + \pi_{\hat{B}} p_{\hat{B},\bar{S}_1}, \tag{140a}$$

$$\pi_{\bar{S}_i} = \pi_{\bar{S}_{i-1}} p_{\bar{S}_{i-1},\bar{S}_i} + \pi_{\hat{S}_{i-1}} p_{\hat{S}_{i-1},\bar{S}_i}, \qquad i \in [2, 2N], \tag{140b}$$

$$\pi_{\bar{G}_1} = \pi_{\bar{S}_{2N}} p_{\bar{S}_{2N},\bar{G}_1} + \pi_{\bar{G}_2} p_{\bar{G}_2,\bar{G}_1} + \pi_{\hat{S}_{2N}} p_{\hat{S}_{2N},\bar{G}_1} + \pi_{\hat{G}_2} p_{\hat{G}_2,\bar{G}_1}, \tag{140c}$$

$$\pi_{\bar{G}_2} = \pi_{\bar{G}_1} p_{\bar{G}_1,\bar{G}_2} + \pi_{\hat{G}_1} p_{\hat{G}_1,\bar{G}_2}, \tag{140d}$$

$$\pi_{\bar{A}_1} = \pi_{\bar{B}} p_{\bar{B},\bar{A}_1} + \pi_{\hat{B}} p_{\hat{B},\bar{A}_1}, \tag{140e}$$

$$\pi_{\bar{A}_2} = \sum_{i=1}^{N} \pi_{\bar{S}_{2i-1}} p_{\bar{S}_{2i-1},\bar{A}_2} + \pi_{\bar{G}_1} p_{\bar{G}_1,\bar{A}_2} + \sum_{i=1}^{N} \pi_{\hat{S}_{2i-1}} p_{\hat{S}_{2i-1},\bar{A}_2} + \pi_{\hat{G}_1} p_{\hat{G}_1,\bar{A}_2}, \tag{140f}$$

$$\pi_{\bar{A}_3} = \pi_{\bar{G}_2} p_{\bar{G}_2,\bar{A}_3} + \pi_{\hat{G}_2} p_{\hat{G}_2,\bar{A}_3}, \tag{140g}$$

$$\pi_{\bar{A}_4} = \sum_{i=1}^{N} \pi_{\bar{S}_{2i}} p_{\bar{S}_{2i},\bar{A}_4} + \sum_{i=1}^{N} \pi_{\hat{S}_{2i}} p_{\hat{S}_{2i},\bar{A}_4}. \tag{140h}$$

For the exit states $C_i, i \in [0, 5]$, such probabilities follow these relationships:

$$\pi_{C_0} = \pi_{V_1} p_{V_1,C_0}, \tag{141a}$$

$$\begin{aligned}\pi_{C_1} = &\sum_{i=1}^{4} \pi_{\bar{A}_i} p_{\bar{A}_i,C_1} + \sum_{i=1}^{4} \pi_{\hat{A}_i} p_{\hat{A}_i,C_1} + \pi_{\bar{B}} p_{\bar{B},C_1} + \pi_{\hat{B}} p_{\hat{B},C_1} \\ &+ \sum_{i=1}^{N} \pi_{\bar{S}_{2i-1}} p_{\bar{S}_{2i-1},C_1} + \sum_{i=1}^{N} \pi_{\hat{S}_{2i-1}} p_{\hat{S}_{2i-1},C_1} + \pi_{\bar{G}_1} p_{\bar{G}_1,C_1} + \pi_{\hat{G}_1} p_{\hat{G}_1,C_1},\end{aligned} \tag{141b}$$

$$\pi_{C_2} = \pi_{\hat{G}_2} p_{\hat{G}_2,C_2}, \tag{141c}$$

$$\pi_{C_3} = \pi_{\bar{G}_2} p_{\bar{G}_2,C_3}, \tag{141d}$$

$$\pi_{C_4} = \sum_{i=1}^{N} \pi_{\hat{S}_{2i}} p_{\hat{S}_{2i},C_4}, \tag{141e}$$

$$\pi_{C_5} = \sum_{i=1}^{N} \pi_{\bar{S}_{2i}} p_{\bar{S}_{2i},C_5}. \tag{141f}$$

Compared to the formulation of the EMC for the anchor CC and the SCell, the one of the internal states of the "deep sleep" cannot be expressed in a compact way similar to the one of Eq. (96) or Eq. (130). Nevertheless, it can be numerically solved in terms of π_{V_1} once the transition probabilities are known. Such transition probabilities depend on the BS packet arrival model, the probability distribution, and their service discipline not only at the SCell, but also at the anchor CC. To differentiate the parameters of the SCell from those

of the anchor, we utilize the following notation for the latter: Λ_1 for the number of packets that arrive in a single subframe, $\lambda_1(k)$ for the PMF of Λ_1, λ_1 for the mean value of Λ_1, X_1 for the service time of a single packet, $b_1(l)$ for the PMF of X_1, and b_1 for the mean value of X_1. We now proceed to characterize the transition probabilities.

4.3.3.1 Transition Probabilities to the Synchronization States

The transition probabilities from V_1 to each of the synchronization states depend directly on the stationary probabilities of the EMC of the anchor CC and on whether a packet arrives at the BS for the SCell during the last subframe of V_1, as shown in Figure 38.

$$p_{V_1,C_0} = [\lambda_2(0)]^{T_{\text{on2}}-1}\left(1-\lambda_2(0)\right)\left(\tilde{\pi}_B + \tilde{\pi}_{G_1} + \sum_{i=1}^{N}\tilde{\pi}_{S_{2i-1}} + \sum_{i=1}^{4}\tilde{\pi}_{A_i}\right), \tag{142a}$$

$$p_{V_1,\hat{B}} = [\lambda_2(0)]^{T_{\text{on2}}}\,\tilde{\pi}_B, \tag{142b}$$

$$p_{V_1,\hat{A}_i} = [\lambda_2(0)]^{T_{\text{on2}}}\,\tilde{\pi}_{A_i}, \qquad i \in [1,4], \tag{142c}$$

$$p_{V_1,\hat{S}_{2i-1}} = [\lambda_2(0)]^{T_{\text{on2}}}\,\tilde{\pi}_{S_{2i-1}}, \qquad i \in [1,N], \tag{142d}$$

$$p_{V_1,\hat{S}_{2i}} = [\lambda_2(0)]^{T_{\text{on2}}-1}\,\tilde{\pi}_{S_{2i}}, \qquad i \in [1,N], \tag{142e}$$

$$p_{V_1,\hat{G}_1} = [\lambda_2(0)]^{T_{\text{on2}}}\,\tilde{\pi}_{G_1}, \tag{142f}$$

$$p_{V_1,\hat{G}_2} = [\lambda_2(0)]^{T_{\text{on2}}-1}\,\tilde{\pi}_{G_2}, \tag{142g}$$

where for every state U, $\tilde{\pi}_U$ is obtained from Eq. (115). The transition probability p_{V_1,C_0} captures the event that no packet arrives for the SCell during the first $T_{\text{on1}} - 1$ subframes of V_1, a packet arrives for the SCell during the last subframe, i.e., subframe T_{on1} of V_1, and the anchor CC is in a non-"sleep" state during subframe $T_{\text{on1}} + 1$. When such event occurs, state C_0 is utilized to immediately notify the user of the need to transition the SCell out of the "deep sleep" state.

The transition probabilities from the synchronization states to the rest of the states in Figure 38 depend on the number of subframes left for the anchor to transition to its next

state, which itself depends on the time instant when the SCell DRX lands in the synchronization state.

4.3.3.2 Transition Probabilities from the Synchronization "On" States

The synchronization "on" states correspond to states $\hat{S}_{2i-1}, i \in [1, N]$, and $\hat{G}_1$. Consider a "sleep" state whose maximum length is v subframes. Denote by P the subframe to be executed at the moment that the transition from V_1 takes place, i.e., there are up to $v-(P-1)$ subframes left until the end of the state. Then, given that $P = i, i \in [1, v]$,

$$\text{Prob}\{\Omega_1|P = i, v\} = \sum_{j=1}^{v-(i-1)-1} q^{j-1}\lambda_2(0)\left[1-\lambda_1(0)\right] = \lambda_2(0)\left[1-\lambda_1(0)\right]\frac{1-q^{v-(i-1)-1}}{1-q}, \tag{143a}$$

$$\text{Prob}\{\Omega_2|P = i, v\} = \sum_{j=1}^{v-(i-1)-1} q^{j-1}\left[1-\lambda_2(0)\right] = \left[1-\lambda_2(0)\right]\frac{1-q^{v-(i-1)-1}}{1-q}, \tag{143b}$$

$$\text{Prob}\{\Omega_3|P = i, v\} = q^{v-(i-1)-1}, \tag{143c}$$

where $q = \lambda_1(0)\lambda_2(0)$. Ω_1, Ω_2, and Ω_3 indicate that the next transition is to a continuous reception state of the anchor CC, to state C_1, and to the next "sleep" state of the anchor CC, respectively. In other words, they capture the required transition probabilities. The factor q^{j-1} is the probability that no packet arrives at the BS for the anchor CC or the SCell during $j-1$ subframes. To obtain the unconditional probabilities of Ω_1, Ω_2, and Ω_3, we need $\text{Prob}\{P = i\}$.

An "on" state can be seen as a group of substates of 1 subframe duration each. In the EMC of such group, the transition probability from every substate to the next one is the probability of no packet arriving in one subframe, i.e., $\lambda_1(0)$. Therefore, the stationary probability π_i of the i-th substate satisfies

$$\pi_i = \pi_{i-1}\lambda_1(0) = \pi_1\left[\lambda_1(0)\right]^{i-1}. \tag{144}$$

Similarly, the stationary probability $\tilde{\pi}_i$ of the i-th state in the semi-Markov Chain is

$$\tilde{\pi}_i = \frac{\pi_i H_i}{\sum\limits_{\forall U'} \pi_{U'} H_{U'}} = \frac{\pi_i}{\sum\limits_{\forall U'} \pi_{U'} H_{U'}} = \frac{\pi_1}{\sum\limits_{\forall U'} \pi_{U'} H_{U'}} \left[\lambda_1(0)\right]^{i-1} = \tilde{\pi}_1 \left[\lambda_1(0)\right]^{i-1}. \tag{145}$$

Since the stationary probability $\tilde{\pi}_{\text{on}}$ of an "on" state, seen as a single state, in the semi-Markov Chain is equivalent to the sum of the stationary probability $\tilde{\pi}_i$ of its substates, then

$$\tilde{\pi}_{\text{on}} = \sum_{i=1}^{v} \tilde{\pi}_i = \sum_{i=1}^{v} \tilde{\pi}_1 \left[\lambda_1(0)\right]^{i-1} = \tilde{\pi}_1 \frac{1 - \left[\lambda_1(0)\right]^v}{1 - \left[\lambda_1(0)\right]} = \frac{\tilde{\pi}_i}{\left[\lambda_1(0)\right]^{i-1}} \frac{1 - \left[\lambda_1(0)\right]^v}{1 - \left[\lambda_1(0)\right]}. \tag{146}$$

Therefore,

$$\tilde{\pi}_i = \left[\lambda_1(0)\right]^{i-1} \frac{1 - \left[\lambda_1(0)\right]}{1 - \left[\lambda_1(0)\right]^v} \tilde{\pi}_{\text{on}}, \tag{147}$$

from which we then obtain that

$$\text{Prob}\left\{P = i|v\right\} = \left[\lambda_1(0)\right]^{i-1} \frac{1 - \left[\lambda_1(0)\right]}{1 - \left[\lambda_1(0)\right]^v}. \tag{148}$$

We can now obtain the unconditional probabilities of Ω_1, Ω_2, and Ω_3 as

$$\text{Prob}\left\{\Omega_j|v\right\} = \sum_{i=1}^{v} \text{Prob}\left\{\Omega_j|P = i, v\right\} \text{Prob}\left\{P = i|v\right\}, \qquad i \in [1, 3], \tag{149}$$

from which we then get

$$\text{Prob}\left\{\Omega_1|v\right\} = \frac{\lambda_2(0) - q}{1 - q} \left[1 - f_1(v)\right], \tag{150a}$$

$$\text{Prob}\left\{\Omega_2|v\right\} = \frac{1 - \lambda_2(0)}{1 - q} \left[1 - f_1(v)\right], \tag{150b}$$

$$\text{Prob}\left\{\Omega_3|v\right\} = f_1(v), \tag{150c}$$

where

$$f_1(x) = \left[\lambda_1(0)\right]^{x-1} \frac{1 - \lambda_1(0)}{1 - \left[\lambda_1(0)\right]^x} \frac{1 - \left[\lambda_2(0)\right]^x}{1 - \lambda_2(0)}, \tag{151}$$

and $q = \lambda_1(0)\lambda_2(0)$. Then, the transition probabilities from the "on" synchronization states of the short DRX are

$$p_{\hat{S}_{2i-1},\bar{A}_2} = \text{Prob}\{\Omega_1 | v = T_{\text{on1}}\} = \frac{\lambda_2(0) - q}{1 - q}\left[1 - f_1\left(T_{\text{on1}}\right)\right], \qquad i \in [1, N], \tag{152a}$$

$$p_{\hat{S}_{2i-1},C_1} = \text{Prob}\{\Omega_2 | v = T_{\text{on1}}\} = \frac{1 - \lambda_2(0)}{1 - q}\left[1 - f_1\left(T_{\text{on1}}\right)\right], \qquad i \in [1, N], \tag{152b}$$

$$p_{\hat{S}_{2i-1},\bar{S}_{2i}} = \text{Prob}\{\Omega_3 | v = T_{\text{on1}}\} = f_1\left(T_{\text{on1}}\right), \qquad i \in [1, N]. \tag{152c}$$

Similarly, the ones of the long DRX are

$$p_{\hat{G}_1,\bar{A}_2} = \frac{\lambda_2(0) - q}{1 - q}\left[1 - f_1\left(T_{\text{on1}}\right)\right], \qquad i \in [1, N], \tag{153a}$$

$$p_{\hat{G}_1,C_1} = \frac{1 - \lambda_2(0)}{1 - q}\left[1 - f_1\left(T_{\text{on1}}\right)\right], \qquad i \in [1, N], \tag{153b}$$

$$p_{\hat{G}_1,\bar{G}_2} = f_1\left(T_{\text{on1}}\right), \qquad i \in [1, N]. \tag{153c}$$

In addition to the transition probabilities, for $i \in [1, v]$, important relationships include

$$\text{Prob}\{L = k | P = i, v\} = \begin{cases} q^{k-1}(1 - q) & : 1 \le k < v - (i - 1) \\ q^{k-1} & : k = v - (i - 1) \end{cases}, \qquad v > 1, \tag{154}$$

$$\text{Prob}\{L = k | P = i, v\} = 1, \qquad k = v = 1, \tag{155}$$

$$E\{L | P = i, v\} = \frac{1 - q^{v-(i-1)}}{1 - q}, \tag{156}$$

where L denotes the number of subframes spent in the "on" synchronization state. From the above, it follows that

$$E\{L | v\} = E\{E\{L | P = i, v\}\} = \frac{1 - q f_1(v)}{1 - q}. \tag{157}$$

Conceptually, $E\{L|v\}$ represents the holding time of an "on" synchronization state.

4.3.3.3 Transition Probabilities from the Synchronization Inactivity Period

The synchronization inactivity period corresponds to state $\hat{B}$. An inactivity period is very similar to an "on" state. The main difference is that a transition to a continuous reception state is possible from the last subframe of the inactivity period, and not of the "on" period. This difference impacts the formulation of the transition probabilities.

Consider that the inactivity period has a maximum length of v subframes. Denote by P the subframe to be executed at the moment that the transition from V_1 takes place, i.e., there are up to $v-(P-1)$ subframes left until the end of the state. Then, given that $P = i, i \in [1, v]$,

$$\text{Prob}\{\Omega_1|P=i,v\} = \sum_{j=1}^{v-(i-1)} q^{j-1}\lambda_2(0)\left[1-\lambda_1(0)\right] = \lambda_2(0)\left[1-\lambda_1(0)\right]\frac{1-q^{v-(i-1)}}{1-q}, \quad (158a)$$

$$\text{Prob}\{\Omega_2|P=i,v\} = \sum_{j=1}^{v-(i-1)} q^{j-1}\left[1-\lambda_2(0)\right] = \left[1-\lambda_2(0)\right]\frac{1-q^{v-(i-1)}}{1-q}, \quad (158b)$$

$$\text{Prob}\{\Omega_3|P=i,v\} = q^{v-(i-1)}, \quad (158c)$$

where $q = \lambda_1(0)\lambda_2(0)$. Ω_1, Ω_2, and Ω_3 indicate that the next transition is to a continuous reception state of the anchor CC, to C_1, and to the next "on" state of the anchor CC, respectively. In other words, they capture the required transition probabilities. The factor q^{j-1} is the probability that no packet arrives at the BS for the anchor CC or the SCell during $j-1$ subframes. To obtain the unconditional probabilities of Ω_1, Ω_2, and Ω_3, we need $\text{Prob}\{P=i\}$. Such probability is found by following a similar analysis as the one for the synchronization "on" state, i.e., it follows Eq. (148). Therefore, the unconditional probabilities of Ω_1, Ω_2, and Ω_3 become

$$\text{Prob}\{\Omega_1|v\} = \frac{\lambda_2(0)-q}{1-q}\left[1-qf_1(v)\right], \quad (159a)$$

$$\text{Prob}\{\Omega_2|v\} = \frac{1-\lambda_2(0)}{1-q}\left[1-qf_1(v)\right], \quad (159b)$$

$$\text{Prob}\{\Omega_3|v\} = qf_1(v), \quad (159c)$$

where $f_1(x)$ is defined in Eq. (151). We can now obtain the transition probabilities:

$$p_{\hat{B},C_1} = \text{Prob}\{\Omega_2|v = T_{\alpha 1}\} = \frac{1-\lambda_2(0)}{1-q}\left[1 - qf_1(T_{\alpha 1})\right], \tag{160a}$$

$$p_{\hat{B},\bar{A}_1} = \text{Prob}\{\Omega_1|v = T_{\alpha 1}\} = \frac{\lambda_2(0)-q}{1-q}\left[1 - qf_1(T_{\alpha 1})\right], \tag{160b}$$

$$p_{\hat{B},\bar{S}_1} = \text{Prob}\{\Omega_3|v = T_{\alpha 1}\} = qf_1(T_{\alpha 1}). \tag{160c}$$

By following a similar analysis as the one for the synchronization "on" state, the holding time of the synchronization inactivity period is found to follow Eq. (157).

4.3.3.4 Transition Probabilities from the Synchronization "Sleep" States

The synchronization "sleep" states correspond to states $\hat{S}_{2i}, i \in [1, N]$, and $\hat{G}_2$. Consider a "sleep" state whose maximum length is v subframes. Denote by P the subframe to be executed at the moment that the transition from V_1 takes place, i.e., there are $v - (P-1)$ subframes left until the end of the state. Then, given that $P = i, i \in [1, v]$,

$$\text{Prob}\{R = 0|P = i, v\} = [\lambda_2(0)]^{v-(i-1)+1}, \tag{161a}$$

$$\begin{aligned}\text{Prob}\{\Omega_1|P = i, v\} &= \left[1 - [\lambda_1(0)]^v\right] \text{Prob}\{R = 0|P = i, v\} \\ &= \left[1 - [\lambda_1(0)]^v\right][\lambda_2(0)]^{v-(i-1)+1},\end{aligned} \tag{161b}$$

$$\text{Prob}\{\Omega_2|P = i, v\} = 1 - \text{Prob}\{R = 0|P = i, v\} = 1 - [\lambda_2(0)]^{v-(i-1)+1}, \tag{161c}$$

$$\text{Prob}\{\Omega_3|P = i, v\} = [\lambda_1(0)]^v \text{Prob}\{R = 0|P = i, v\} = [\lambda_1(0)]^v [\lambda_2(0)]^{v-(i-1)+1}, \tag{161d}$$

where R is the number of packets that arrive at the BS for the SCell by the end of the state. Ω_1, Ω_2, and Ω_3 indicate that the next transition is to a continuous reception state of the anchor CC, to a state C_i, and to the next "on" state of the anchor CC, respectively. For a "sleep" state, P is a discrete and uniformly distributed random variable in the range $[1, v]$.

It follows that

$$\begin{aligned}\text{Prob}\{R=0|v\} &= \sum_{i=1}^{v} \text{Prob}\{R=0|P=i,v\}\,\text{Prob}\{P=i|v\} \\ &= \frac{[\lambda_2(0)]^2}{1-\lambda_2(0)}\frac{1-[\lambda_2(0)]^v}{v}. \end{aligned} \tag{162}$$

Let

$$f_2(x) = \frac{[\lambda_2(0)]^2}{1-\lambda_2(0)}\frac{1-[\lambda_2(0)]^x}{x}. \tag{163}$$

Then,

$$\text{Prob}\{\Omega_1|v\} = \left[1-[\lambda_1(0)]^v\right] f_2(v), \tag{164a}$$

$$\text{Prob}\{\Omega_2|v\} = 1-f_2(v), \tag{164b}$$

$$\text{Prob}\{\Omega_3|v\} = [\lambda_1(0)]^v f_2(v). \tag{164c}$$

Therefore, the transition probabilities from the "sleep" periods are

$$p_{\hat{S}_{2i},C_4} = \text{Prob}\left\{\Omega_2|v=T_{\beta 1}-T_{\text{on}1}\right\} = 1-f_2\left(T_{\beta 1}-T_{\text{on}1}\right), \qquad i\in[1,N], \tag{165a}$$

$$\begin{aligned} p_{\hat{S}_{2i},\bar{S}_{2i+1}} &= \text{Prob}\left\{\Omega_3|v=T_{\beta 1}-T_{\text{on}1}\right\} \\ &= [\lambda_1(0)]^{T_{\beta 1}-T_{\text{on}1}+1} f_2\left(T_{\beta 1}-T_{\text{on}1}\right), \qquad i\in[1,N-1], \end{aligned} \tag{165b}$$

$$\begin{aligned} p_{\hat{S}_{2N},\bar{G}_1} &= \text{Prob}\left\{\Omega_3|v=T_{\beta 1}-T_{\text{on}1}\right\} \\ &= [\lambda_1(0)]^{T_{\beta 1}-T_{\text{on}1}+1} f_2\left(T_{\beta 1}-T_{\text{on}1}\right), \end{aligned} \tag{165c}$$

$$\begin{aligned} p_{\hat{S}_{2i},\bar{A}_4} &= \text{Prob}\left\{\Omega_1|v=T_{\beta 1}-T_{\text{on}1}\right\} \\ &= \left[1-[\lambda_1(0)]^{T_{\beta 1}-T_{\text{on}1}+1}\right] f_2\left(T_{\beta 1}-T_{\text{on}1}\right), \qquad i\in[1,N], \end{aligned} \tag{165d}$$

$$p_{\hat{G}_2,C_2} = \text{Prob}\left\{\Omega_2|v=T_{\gamma 1}-T_{\text{on}1}\right\} = 1-f_2\left(T_{\gamma 1}-T_{\text{on}1}\right), \tag{165e}$$

$$\begin{aligned} p_{\hat{G}_2,\bar{G}_1} &= \text{Prob}\left\{\Omega_3|v=T_{\gamma 1}-T_{\text{on}1}\right\} \\ &= [\lambda_1(0)]^{T_{\gamma 1}-T_{\text{on}1}+1} f_2\left(T_{\gamma 1}-T_{\text{on}1}\right), \end{aligned} \tag{165f}$$

$$p_{\hat{G}_2,\bar{A}_3} = \text{Prob}\left\{\Omega_2 1 | v = T_{\gamma 1} - T_{\text{on}1}\right\}$$
$$= \left[1 - [\lambda_1(0)]^{T_{\gamma 1} - T_{\text{on}1} + 1}\right] f_2\left(T_{\gamma 1} - T_{\text{on}1}\right). \tag{165g}$$

In addition to the transition probabilities, important relationships also include

$$E\{R|P = i, v\} = [v - (i-1) + 1]\,\lambda_2, \tag{166}$$
$$E\{R|v\} = E\{E\{R|P = i, v\}\} = \left[\frac{v+1}{2} + 1\right]\lambda_2, \tag{167}$$
$$\text{Prob}\{L = i|v\} = \text{Prob}\{P = v - (i-1)|v\} = \frac{1}{v}, \qquad i \in [1, v], \tag{168}$$

where L denotes the number of subframes left to be executed until the end of the "sleep" state. Since L is also uniformly distributed over $[1, v]$, we have that

$$E\{L|v\} = \frac{v+1}{2}. \tag{169}$$

Let $\check{R}$ denote R conditioned on being greater than zero, i.e., $\check{R}$ is the number of packets in the BS for the SCell given that at least one such packet arrived. Then,

$$\text{Prob}\left\{\check{R} = k|v\right\} = \begin{cases} \frac{\text{Prob}\{R=k|v\}}{1-\text{Prob}\{R=0|v\}} & : k > 0 \\ 0 & : k = 0 \end{cases}, \tag{170}$$
$$E\{\check{R}|v\} = \frac{E\{R|v\}}{1 - \text{Prob}\{R=0|v\}} = \frac{1 + \frac{v+1}{2}}{1 - f_2(v)}\lambda_2. \tag{171}$$

Conceptually, $\check{R}$ represents the number of buffered packets for the SCell if the BS determines that it should trigger the activation of the SCell at the end of a "sleep" synchronization state, and $E\{L|v\}$ represents the holding time of a "sleep" synchronization state.

4.3.3.5 Transition Probabilities from the Synchronization Continuous Reception States

Compared to the previously described synchronization states, a synchronization continuous reception state does not have a maximum duration that applies to every instance of such

state.

Let P_{-} be a random variable corresponding to the number of subframes left until the end of the synchronization continuous reception state at the moment that the transition from V_1 takes place, i.e., if the UE were not tracking such state for the presence of a PDCCH for the SCell, such state would have executed for P_{-} more subframes. Then, it follows that

$$\text{Prob}\{\Omega_2 | P_{-} = m\} = 1 - [\lambda_2(0)]^m, \tag{172a}$$

$$\text{Prob}\{\Omega_3 | P_{-} = m\} = [\lambda_2(0)]^m, \tag{172b}$$

where Ω_2 and Ω_3 indicate that the next transition is to C_1, and to the next inactivity period, respectively. To obtain the unconditional probabilities of Ω_2 and Ω_3, we need $\text{Prob}\{P_{-} = m\}$. In contrast to the transition probabilities discussed in the previous section, there is no explicit formulation for $\text{Prob}\{P_{-} = m\}$; nevertheless, we can obtain one for its probability-generating function (PGF). Such approach is useful because

$$\text{Prob}\{\Omega_2\} = \sum_{i=1}^{\infty} \text{Prob}\{\Omega_2 | P_{-} = m\} \, \text{Prob}\{P_{-} = m\} = 1 - \mathcal{Z}_{P_{-}}(z)\Big|_{z=\lambda_2(0)}, \tag{173a}$$

$$\text{Prob}\{\Omega_3\} = \sum_{i=1}^{\infty} \text{Prob}\{\Omega_3 | P_{-} = m\} \, \text{Prob}\{P_{-} = m\} = \mathcal{Z}_{P_{-}}(z)\Big|_{z=\lambda_2(0)}, \tag{173b}$$

where

$$\mathcal{Z}_{P_{-}}(z) = \sum_{m=1}^{\infty} z^m \, \text{Prob}\{P_{-} = m\}, \tag{174}$$

i.e., $\mathcal{Z}_{P_{-}}(z)$ represents the PGF of P_{-}. We now describe how to obtain an expression for such PGF.

Let P be a random variable corresponding to the number of subframes of the continuous reception state of interest. Let P_{+} be a random variable corresponding to the number of subframes of the continuous reception state landed after transitioning from V_1. An instance of a continuous reception state lasting j subframes has a chance of including the landing subframe with a probability proportional to j. Therefore, the probability that P_{+} has j

subframes is given by

$$\text{Prob}\{P_+ = j\} = \frac{j\,\text{Prob}\{P = j\}}{E\{P\}}. \tag{175}$$

The position of the landed subframe within the synchronization continuous reception state is uniformly distributed over the length of the state. Hence, the probability that there are k remaining subframes in the synchronization continuous reception state is

$$\text{Prob}\{P_- = k | P_+ = j\} = \frac{1}{j}, \qquad j \geq 1, \quad k = 1, 2, \ldots, j. \tag{176}$$

So,

$$\text{Prob}\{P_- = k\} = \sum_{j=1}^{\infty} \text{Prob}\{P_- = k | P_+ = j\}\,\text{Prob}\{P_+ = j\} = \frac{\text{Prob}\{P \geq k\}}{E\{P\}}, \tag{177}$$

then, the PGF of P_- becomes

$$\mathcal{Z}_{P_-}(z) = \frac{z}{1-z}\frac{1 - \mathcal{Z}_P(z)}{E\{P\}}, \tag{178}$$

where $\mathcal{Z}_P(z)$ is the PGF of P, i.e., the PGF of the length (in subframes) of a continuous reception state. If such state started with $\check{R}$ subframes, then

$$P = \sum_{i=1}^{\check{R}} P_*, \tag{179}$$

where P_* is a random variable corresponding to the length of a continuous reception state caused by one packet in the buffer, i.e., the length of a busy period caused by one packet, in queuing theory terminology. Therefore, the PGF of P becomes

$$\mathcal{Z}_P(z) = \mathcal{Z}_{\check{R}}\Big(\mathcal{Z}_{P_*}(z)\Big), \tag{180}$$

where $\mathcal{Z}_{\check{R}}(z)$ is the PGF of $\check{R}$. Since the PMF of $\check{R}(z)$ follows (170), its PGF is

$$\mathcal{Z}_{\check{R}}(z) = \frac{\mathcal{Z}_R(z) - \text{Prob}\{R = 0\}}{1 - \text{Prob}\{R = 0\}}, \tag{181}$$

where R is the number of packets buffered during the state that preceded the continuous reception state. If such state buffered packets during v subframes, then we have that $\text{Prob}\{R = 0\} = [\lambda_1(0)]^v$ and $\mathcal{Z}_R(z) = [\mathcal{Z}_{\Lambda_1}(z)]^v$. Therefore,

$$\mathcal{Z}_{\check{R}}(z) = \frac{[\mathcal{Z}_{\Lambda_1}(z)]^v - [\lambda_1(0)]^v}{1 - [\lambda_1(0)]^v}, \tag{182}$$

and plugging into Eq. (180), we get

$$\mathcal{Z}_P(z) = \frac{\left[\mathcal{Z}_{\Lambda_1}\left(\mathcal{Z}_{P_*}(z)\right)\right]^v - [\lambda_1(0)]^v}{1 - [\lambda_1(0)]^v}. \tag{183}$$

The expected value of P can be found from

$$E\{P\} = \frac{d}{dz}\mathcal{Z}_P(z)\Big|_{z=1} = \frac{v\lambda_1}{1 - [\lambda_1(0)]^v} E\{P_*\}, \tag{184}$$

where $E\{P_*\}$ is obtained from Eq. (109) by considering that the initial number of packets in the buffer is 1, i.e.,

$$E\{P_*\} = \frac{b_1}{1 - \rho_1}, \tag{185}$$

where $\rho_1 = \lambda_1 b_1$. Therefore, by plugging Eq. (185) into Eq. (184), we have

$$E\{P\} = \frac{\rho_1}{1 - \rho_1} \frac{v}{1 - [\lambda_1(0)]^v}. \tag{186}$$

By plugging Eq. (183) and Eq. (186) into Eq. (178), we obtain

$$\mathcal{Z}_{P_-}(z) = \frac{z}{1-z}\frac{1-\rho_1}{\rho_1}\frac{1-\left[\mathcal{Z}_{\Lambda_1}\left(\mathcal{Z}_{P_*}(z)\right)\right]^v}{v}. \tag{187}$$

From Eq. (173a) and Eq. (173b), we can now obtain the probabilities of Ω_1 and Ω_2:

$$\text{Prob}\{\Omega_2\} = 1 - f_3(v), \tag{188a}$$

$$\text{Prob}\{\Omega_3\} = f_3(v), \tag{188b}$$

where

$$f_3(x) = \frac{1-\rho_1}{\rho_1}\frac{\lambda_2(0)}{1-\lambda_2(0)}\left.\frac{1-\left[\mathcal{Z}_{\Lambda_1}\left(\mathcal{Z}_{P_*}(z)\right)\right]^x}{x}\right|_{z=\lambda_2(0)}, \tag{189}$$

and $\mathcal{Z}_{P_*}(z)$ is obtained from [82] as the PGF of the length of a busy period triggered by the arrival of one packet:

$$\mathcal{Z}_{P_*}(u) = \sum_{n=1}^{\infty}\frac{u^n}{n!}\frac{d^{n-1}}{dz^{n-1}}\left\{\left[\frac{d}{dz}\mathcal{Z}_{X_1}(z)\right]\left[\mathcal{Z}_{\Lambda_1}\left(\mathcal{Z}_{X_1}(z)\right)\right]^n\right\}\Bigg|_{z=0}. \tag{190}$$

Then, the transition probabilities from the continuous reception periods can be obtained as follows:

$$p_{\hat{A}_1,\bar{B}} = p_{\hat{A}_2,\bar{B}} = f_3(1), \tag{191a}$$

$$p_{\hat{A}_1,C_1} = p_{\hat{A}_2,C_1} = 1 - f_3(1), \tag{191b}$$

$$p_{\hat{A}_4,\bar{B}} = f_3(T_{\beta 1} - T_{\text{on}1} + 1), \tag{191c}$$

$$p_{\hat{A}_4,C_1} = 1 - f_3(T_{\beta 1} - T_{\text{on}1} + 1), \tag{191d}$$

$$p_{\hat{A}_3,\bar{B}} = f_3(T_{\gamma 1} - T_{\text{on}1} + 1), \tag{191e}$$

$$p_{\hat{A}_3,C_1} = 1 - f_3(T_{\gamma 1} - T_{\text{on}1} + 1). \tag{191f}$$

In addition to the transition probabilities, important relationships include

$$\text{Prob}\{L = k|P_- = m\} = \begin{cases} [\lambda_2(0)]^{k-1}[1-\lambda_2(0)] & : 1 \le k < m \\ [\lambda_2(0)]^{k-1} & : k = m \end{cases}, \qquad m > 1, \tag{192}$$

$$\text{Prob}\{L = k|P_- = m\} = 1, \qquad m = k = 1, \tag{193}$$

$$E\{L|P_- = m\} = \frac{1-[\lambda_2(0)]^m}{1-\lambda_2(0)}, \tag{194}$$

where L denotes the number of subframes spent in the synchronization continuous reception state. From the above, it follows that

$$E\{L\} = E\{E\{L|P_- = m\}\} = \frac{1-f_3(v)}{1-\lambda_2(0)}. \tag{195}$$

Conceptually, $E\{L\}$ represents the holding time of a synchronization continuous reception state.

4.3.3.6 Transition Probabilities for Non-Synchronization States

For the non-synchronization states, we use most of the expressions previously developed for the synchronization states. We can do so because a non-synchronization state is equivalent to a synchronization state that always starts at subframe 1 of the original state. Particularly, for the "on," inactivity, and "sleep" states, we only need to set $P = 1$ in Eq. (143), Eq. (158), Eq. (161), respectively. Then, we obtain

$$p_{\bar{B},\bar{S}_1} = q^{T_{\alpha 1}}, \tag{196a}$$

$$p_{\bar{B},\bar{A}_1} = [\lambda_2(0) - q]\frac{1-q^{T_{\alpha 1}}}{1-q}, \tag{196b}$$

$$p_{\bar{B},C_1} = [1-\lambda_2(0)]\frac{1-q^{T_{\alpha 1}}}{1-q}, \tag{196c}$$

$$p_{\bar{S}_{2i-1},\bar{S}_{2i}} = q^{T_{\text{on}1}-1}, \qquad i \in [1,N], \tag{196d}$$

$$p_{\bar{S}_{2i-1},\bar{A}_2} = [\lambda_2(0) - q]\frac{1-q^{T_{\text{on}1}-1}}{1-q}, \qquad i \in [1,N], \tag{196e}$$

$$p_{\bar{S}_{2i-1},C_1} = [1-\lambda_2(0)]\,\frac{1-q^{T_{\text{on1}}-1}}{1-q}, \qquad i \in [1,N], \tag{196f}$$

$$p_{\bar{S}_{2N},\bar{G}_1} = q^{T_{\beta 1}-T_{\text{on1}}+1}, \tag{196g}$$

$$p_{\bar{S}_{2i},\bar{S}_{2i+1}} = q^{T_{\beta 1}-T_{\text{on1}}+1}, \qquad i \in [1,N-1], \tag{196h}$$

$$p_{\bar{S}_{2i},\bar{A}_4} = [\lambda_2(0)]^{T_{\beta 1}-T_{\text{on1}}+1} - q^{T_{\beta 1}-T_{\text{on1}}+1}, \qquad i \in [1,N], \tag{196i}$$

$$p_{\bar{S}_{2i},C_5} = 1-[\lambda_2(0)]^{T_{\beta 1}-T_{\text{on1}}+1}, \qquad i \in [1,N], \tag{196j}$$

$$p_{\bar{G}_2,\bar{G}_1} = q^{T_{\gamma 1}-T_{\text{on1}}+1}, \tag{196k}$$

$$p_{\bar{G}_2,\bar{A}_3} = [\lambda_2(0)]^{T_{\gamma 1}-T_{\text{on1}}+1} - q^{T_{\gamma 1}-T_{\text{on1}}+1}, \tag{196l}$$

$$p_{\bar{G}_2,C_3} = 1-[\lambda_2(0)]^{T_{\gamma 1}-T_{\text{on1}}+1}. \tag{196m}$$

For the continuous reception states, we use the expressions in Eq. (173) and replace the PGF of P_- with the PGF of P (Eq. (183)), leading to

$$p_{\bar{A}_1,C_1} = p_{\bar{A}_2,C_1} = f_4(1), \tag{197a}$$

$$p_{\bar{A}_1,\bar{B}} = p_{\bar{A}_2,\bar{B}} = 1-f_4(1), \tag{197b}$$

$$p_{\bar{A}_3,C_1} = f_4(T_{\gamma 1}-T_{\text{on1}}+1), \tag{197c}$$

$$p_{\bar{A}_3,\bar{B}} = 1-f_4(T_{\gamma 1}-T_{\text{on1}}+1), \tag{197d}$$

$$p_{\bar{A}_4,C_1} = f_4(T_{\beta 1}-T_{\text{on1}}+1), \tag{197e}$$

$$p_{\bar{A}_4,\bar{B}} = 1-f_4(T_{\beta 1}-T_{\text{on1}}+1), \tag{197f}$$

where

$$f_4(x) = \left.\frac{1-\left[\mathcal{Z}_{\Lambda_1}\left(\mathcal{Z}_{P_*}(z)\right)\right]^x}{1-[\lambda_1(0)]^x}\right|_{z=\lambda_2(0)}. \tag{198}$$

At this point, all the transition probabilities have been defined, and the stationary probabilities of the EMC can be evaluated. In addition to the stationary probabilities of the EMC, the holding time of each state is required to compute the energy savings. In Section 4.3.4,

we describe how the holding time is obtained.

4.3.3.7 Transition Probabilities for Exit States

States $C_i, i \in [0, 5]$, correspond to the exit states, i.e., the states that represent the exit from the "deep sleep" state. Their transition probabilities are

$$p_{C_i,F_{i+4}} = 1, \qquad i \in [0, 5]. \tag{199}$$

4.3.4 Holding Time

4.3.4.1 "Deep Sleep" Internal States

For the synchronization states inside the "deep sleep" state, the holding time was discussed jointly with their transition probabilities in Sections 4.3.3.2-4.3.3.4. Here we summarize the expressions for those holding times H:

$$H_{\hat{S}_{2i}} = \frac{T_{\beta 1} - T_{\text{on}1} + 1}{2}, \qquad i \in [1, N], \tag{200a}$$

$$H_{\hat{G}_2} = \frac{T_{\gamma 1} - T_{\text{on}1} + 1}{2}, \tag{200b}$$

$$H_{\hat{S}_{2i-1}} = \frac{1 - q f_1\left(T_{\text{on}1}\right)}{1 - q}, \qquad i \in [1, N], \tag{200c}$$

$$H_{\hat{G}_1} = \frac{1 - q f_1\left(T_{\text{on}1}\right)}{1 - q}, \tag{200d}$$

$$H_{\hat{B}} = \frac{1 - q f_1\left(T_{\alpha 1}\right)}{1 - q}, \tag{200e}$$

$$H_{\hat{A}_1} = H_{\hat{A}_2} = \frac{1 - f_3(1)}{1 - \lambda_2(0)}, \tag{200f}$$

$$H_{\hat{A}_3} = \frac{1 - f_3(T_{\gamma 1} - T_{\text{on}1} + 1)}{1 - \lambda_2(0)}, \tag{200g}$$

$$H_{\hat{A}_4} = \frac{1 - f_3(T_{\beta 1} - T_{\text{on}1} + 1)}{1 - \lambda_2(0)}, \tag{200h}$$

where $f_1(x)$ is defined in Eq. (151), and $f_3(x)$ is defined in Eq. (189).

For the non-synchronization states inside the "deep sleep" state, the holding time is obtained following a similar approach as the one for the transition probabilities in Section 4.3.3.6. Particularly, for the "on" and inactivity states, we only need to set $P = 1$ in

Eq. (156). For the "sleep" and exit states $C_i, i \in [0,5]$, the holding time is a deterministic value:

$$H_{\bar{B}} = \frac{1-q^{T_{\alpha 1}}}{1-q}, \tag{201a}$$

$$H_{\bar{S}_{2i}} = T_{\beta 1} - T_{\text{on}1}, \qquad i \in [1,N], \tag{201b}$$

$$H_{\bar{S}_{2i-1}} = \frac{1-q^{T_{\text{on}1}}}{1-q}, \qquad i \in [1,N], \tag{201c}$$

$$H_{\bar{G}_1} = \frac{1-q^{T_{\text{on}1}}}{1-q}, \tag{201d}$$

$$H_{\bar{G}_2} = T_{\gamma 1} - T_{\text{on}1}, \tag{201e}$$

$$H_{C_i} = 1, \qquad i \in [0,5]. \tag{201f}$$

For the non-synchronization continuous reception states inside the "deep sleep" state, we utilize the expression in Eq. (194) and replace the PGF of P_- with the PGF of P (Eq. (183)) when calculating the holding time, leading to

$$H_{\bar{A}_1} = H_{\bar{A}_2} = \frac{f_4(1)}{1-\lambda_2(0)}, \tag{202a}$$

$$H_{\bar{A}_3} = \frac{f_4\left(T_{\gamma 1} - T_{\text{on}1} + 1\right)}{1-\lambda_2(0)}, \tag{202b}$$

$$H_{\bar{A}_4} = \frac{f_4\left(T_{\beta 1} - T_{\text{on}1} + 1\right)}{1-\lambda_2(0)}, \tag{202c}$$

where $f_4(x)$ is defined in Eq. (198).

4.3.4.2 SCell

For the holding time of the SCell states (Figure 36), we can directly apply the expressions developed in Section 4.2.3 by adjusting for the parameters of the SCell:

$$H_R = \frac{1 - [\lambda_2(0)]^{T_{\alpha 2}}}{1 - \lambda_2(0)}, \tag{203a}$$

$$H_{Y_{2i-1}} = \frac{1 - [\lambda_2(0)]^{T_{\text{on}2}}}{1 - \lambda_2(0)}, \qquad i \in [1, N], \tag{203b}$$

$$H_{Y_{2i}} = T_{\beta 2} - T_{\text{on}2}, \qquad i \in [1, N], \tag{203c}$$

$$H_{V_1} = \frac{1 - [\lambda_2(0)]^{T_{\text{on}2}}}{1 - \lambda_2(0)}, \tag{203d}$$

$$H_{F_1} = H_{F_2} = \frac{\rho_2}{1 - \rho_2} \frac{1}{1 - \lambda_2(0)}, \tag{203e}$$

$$H_{F_3} = \frac{\rho_2}{1 - \rho_2} \frac{T_{\beta 2} - T_{\text{on}2} + 1}{1 - [\lambda_2(0)]^{T_{\beta 2} - T_{\text{on}2} + 1}}. \tag{203f}$$

The previous expressions account for the inactivity period, short DRX cycles, state V_1, and continuous reception states $F_i, i \in [1, 3]$. The last three states are related to the arrival of packets outside the "deep sleep". On the other hand, states $F_i, i \in [4, 9]$, are related to the arrival of packets within the "deep sleep," and their holding times are now analyzed.

From Eq. (110), we have that the holding time of a continuous reception state depends on the expected value of the number of packets in the BS buffer at the moment that the state starts. For states $F_i, i \in [4, 9]$, such expected value $E\{R_{F_i}\}$ is the sum of

- the expected value of the number of packets received during C_{i-4}, i.e., λ_2 packets, and
- the expected value $E\{R_{C_{i-4}}\}$ of the number of packets in the BS buffer when C_{i-4} started,

i.e.,

$$E\{R_{F_i}\} = \lambda_2 + E\{R_{C_{i-4}}\} \qquad i \in [4, 9]. \tag{204}$$

Then, applying Eq. (110), we have that

$$\begin{aligned} H_{F_i} &= E\{R_{F_i}\}\frac{b_2}{1-\rho_2} \\ &= \frac{\rho_2}{1-\rho_2}\left[1+\frac{E\{R_{C_{i-4}}\}}{\lambda_2}\right], \qquad i \in [4,9]. \end{aligned} \tag{205}$$

$E\{R_{C_{i-4}}\}$ can be computed from Eq. (112) for $i \in \{4,5,7,9\}$, i.e., for those states C_i reached from any state except the synchronization "sleep" states. The reason is that the buffering time that precedes states $C_i, i \in \{0,1,3,5\}$, is a deterministic value. So,

$$E\{R_{C_0}\} = E\{R_{C_1}\} = \lambda_2 \frac{1}{1-\lambda_2(0)}, \tag{206a}$$

$$E\{R_{C_3}\} = \lambda_2 \frac{T_{\gamma 1} - T_{\text{on}1} + 1}{1-[\lambda_2(0)]^{T_{\gamma 1}-T_{\text{on}1}+1}}, \tag{206b}$$

$$E\{R_{C_5}\} = \lambda_2 \frac{T_{\beta 1} - T_{\text{on}1} + 1}{1-[\lambda_2(0)]^{T_{\beta 1}-T_{\text{on}1}+1}}. \tag{206c}$$

For $i \in \{2,4\}$, $E\{R_{C_i}\}$ is determined by Eq. (171), i.e., by the expected number of packets at the end of the preceding "sleep" state, given that at least one such packet arrived. So,

$$E\{R_{C_2}\} = \lambda_2 f_5\left(T_{\gamma 1} - T_{\text{on}1}\right), \tag{207a}$$

$$E\{R_{C_4}\} = \lambda_2 f_5\left(T_{\beta 1} - T_{\text{on}1}\right), \tag{207b}$$

where

$$f_5(x) = \frac{1+\frac{x+1}{2}}{1-f_2(x)}, \tag{208}$$

and $f_2(x)$ is defined in Eq. (163). Having $E\{R_{C_i}\}, i \in [0,5]$, we can plug it into Eq. (205) and find the remaining holding times:

$$H_{F_4} = H_{F_5} = \frac{\rho_2}{1-\rho_2}\left[1+\frac{1}{1-\lambda_2(0)}\right], \tag{209a}$$

$$H_{F_6} = \frac{\rho_2}{1-\rho_2}\left[1+f_5\left(T_{\gamma 1} - T_{\text{on}1}\right)\right], \tag{209b}$$

$$H_{F_7} = \frac{\rho_2}{1-\rho_2}\left[1 + \frac{T_{\gamma 1} - T_{\text{on}1} + 1}{1 - [\lambda_2(0)]^{T_{\gamma 1} - T_{\text{on}1} + 1}}\right], \tag{209c}$$

$$H_{F_8} = \frac{\rho_2}{1-\rho_2}\left[1 + f_5\left(T_{\beta 1} - T_{\text{on}1}\right)\right], \tag{209d}$$

$$H_{F_9} = \frac{\rho_2}{1-\rho_2}\left[1 + \frac{T_{\beta 1} - T_{\text{on}1} + 1}{1 - [\lambda_2(0)]^{T_{\beta 1} - T_{\text{on}1} + 1}}\right]. \tag{209e}$$

4.3.5 Performance Metrics

The main performance metrics associated with any DRX scheme are the amount of energy saved and the packet delay, also known as waiting time in queuing theory terminology. Since the metrics for the anchor CC correspond to the ones already analyzed and evaluated in Section 4.2, here we focus on evaluating the performance metrics for the SCell.

4.3.5.1 Energy Savings

The amount of energy saved is defined as the total amount of time spent in the "sleep" and "deep sleep" states. This value can be obtained from the stationary probabilities of the semi-Markov Chain, which we derive from the stationary probabilities of the EMC in Eq. (138), Eq. (139), Eq. (140), and Eq. (141), obtained in Sections 4.3.2 and 4.3.3.

For any state U, its stationary probability $\tilde{\pi}_U$ in the semi-Markov Chain is defined by Eq. (115). Then, the energy savings $\tau_{\beta 2}$ and τ_{ds} provided by the short DRX cycles and the "deep sleep" state, respectively, are

$$\tau_{\beta 2} = \sum_{i=1}^{M} \tilde{\pi}_{Y_{2i}}, \qquad \tau_{\text{ds}} = \sum_{\forall U \in V_2}^{M} \tilde{\pi}_U. \tag{210}$$

Replacing the expressions for the stationary probabilities of the semi-Markov Chain, $\tau_{\beta 2}$ and τ_{ds} become

$$\tau_{\beta 2} = \frac{T_{\beta 2} - T_{\text{on}2}}{\Psi_2} \frac{1 - [\lambda_2(0)]^{MT_{\beta 2}}}{1 - [\lambda_2(0)]^{T_{\beta 2}}} \left[\lambda_2(0)\right]^{T_{\text{on}2}-1}, \tag{211a}$$

$$\tau_{\text{ds}} = \frac{[\lambda_2(0)]^{MT_{\beta 2}}}{\Psi_2} \sum_{U \in V_2} \frac{\pi_U}{\pi_{V_1}} H_U, \tag{211b}$$

where

$$\Psi_2 = \frac{1}{[\lambda_2(0)]^{T_{\alpha 2}}}\left[H_R + \left(1 - [\lambda_2(0)]^{T_{\alpha 2}}\right) H_{F_1}\right] + \left[1 - [\lambda_2(0)]^{T_{\text{on}2}-1}\right][\lambda_2(0)]^{MT_{\beta 2}} H_{F_2}$$
$$+ \left[1 - [\lambda_2(0)]^{MT_{\beta 2}}\right] H_{F_3} + [\lambda_2(0)]^{MT_{\beta 2}}\left[H_{Y_1} + \sum_{i=4}^{9} \frac{\pi_{C_{i-4}}}{\pi_{V_1}} H_{F_i} + \sum_{\forall U \in V_2} \frac{\pi_U}{\pi_{V_1}} H_U\right]$$
$$+ \frac{1 - [\lambda_2(0)]^{MT_{\beta 2}}}{1 - [\lambda_2(0)]^{T_{\beta 2}}}\left[H_{Y_1} + [\lambda_2(0)]^{T_{\text{on}2}-1} H_{Y_2} + \left[1 - [\lambda_2(0)]^{T_{\text{on}2}-1}\right][H_{F_2} - H_{F_3}]\right]. \quad (212)$$

Then, the total energy savings τ_2 become

$$\tau_2 = \tau_{\beta 2} + \tau_{\text{ds}}. \quad (213)$$

4.3.5.2 Delay

To calculate the expected value $E\{\Gamma_2\}$ of the packet delay in the SCell, we need to compute (a) the expected value of the delay Υ_i experienced by the packets sent in $F_i, i \in [1, 9]$, and (b) the probability of a packet to be sent in each of such states. As discussed in Section 4.2.4.2, we apply the results from queuing theory that establish the expected value $E\{\Upsilon\}$ of the packet waiting time in a system with vacation [82]. In such context,

$$E\{\Upsilon\} = \frac{[\lambda_2]^2 E\{[X_2]^2\} + b_2 E\{[\Lambda_2]^2\} - \rho_2(\lambda_2 + 1)}{2\lambda_2(1 - \rho_2)} + \frac{E\{v(v-1)\}}{2E\{v\}}, \quad (214)$$

where v is the length of the vacation, the first term represents the waiting time in a system without vacation, and the second term represents the residual life of the vacation time. In the context of DRX, v corresponds to the amount of time the BS buffers packets before entering a continuous reception state. Therefore, v is a deterministic value equal to 1 for F_1 and F_2 and $T_{\beta 2} - T_{\text{on}2} + 1$ for F_3. For $F_i, i \in [4, 9]$, v is the sum of the amount spent in state C_{i-4}, i.e., 1 subframe, and the amount of buffering time preceding that state. In particular, v is a deterministic value equal to 2 for F_4 and F_5, $T_{\gamma 1} + T_{\text{on}1} + 2$ for F_7, and $T_{\beta 1} - T_{\text{on}1} + 2$ for F_9. On the other hand, v is a discrete uniformly distributed random variable in the range

$[3, T_{\gamma 1} - T_{\text{on}1} + 2]$ for F_6 and in the range $[3, T_{\beta 1} - T_{\text{on}1} + 2]$ for F_8. It then follows that

$$E\{\Upsilon_1\} = E\{\Upsilon_2\} = \frac{[\lambda_2]^2 \, E\{[X_2]^2\} + b_2 E\{[\Lambda_2]^2\} - \rho_2(\lambda_2 + 1)}{2\lambda_2(1 - \rho_2)}, \tag{215a}$$

$$E\{\Upsilon_3\} = E\{\Upsilon_1\} + \frac{T_{\beta 2} - T_{\text{on}2}}{2}, \tag{215b}$$

$$E\{\Upsilon_4\} = E\{\Upsilon_5\} = E\{\Upsilon_1\} + \frac{1}{2}, \tag{215c}$$

$$E\{\Upsilon_6\} = E\{\Upsilon_1\} + \frac{1}{3}\frac{\left(T_{\gamma 1} - T_{\text{on}1}\right)^2 + 6\left(T_{\gamma 1} - T_{\text{on}1}\right) + 11}{\left(T_{\gamma 1} - T_{\text{on}1}\right) + 5}, \tag{215d}$$

$$E\{\Upsilon_7\} = E\{\Upsilon_1\} + \frac{T_{\gamma 1} + T_{\text{on}1} + 1}{2}, \tag{215e}$$

$$E\{\Upsilon_8\} = E\{\Upsilon_1\} + \frac{1}{3}\frac{\left(T_{\beta 1} - T_{\text{on}1}\right)^2 + 6\left(T_{\beta 1} - T_{\text{on}1}\right) + 11}{\left(T_{\beta 1} - T_{\text{on}1}\right) + 5}, \tag{215f}$$

$$E\{\Upsilon_9\} = E\{\Upsilon_1\} + \frac{T_{\beta 1} + T_{\text{on}1} + 1}{2}. \tag{215g}$$

We now proceed to compute the probability of a packet being sent from state F_i. From Eq. (123), we have that such probability is

$$\text{Prob}\{\Phi = F_i\} = \frac{\pi_{F_i} H_{F_i}}{\sum_{j=1}^{9} \pi_{F_j} H_{F_j}}, \tag{216}$$

where Φ denotes the state from which the packet is sent. We can now see that

$$\begin{aligned} E\{\Gamma_2\} = E\{E\{\Gamma_2|\Phi\}\} &= \sum_{i=1}^{9} E\{\Gamma_2|\Phi = F_i\} \, \text{Prob}\{\Phi = F_i\} \\ &= \sum_{i=1}^{9} E\{\Upsilon_i\} \, \text{Prob}\{\Phi = F_i\}. \end{aligned} \tag{217}$$

After further simplification, the above expression becomes

$$
\begin{aligned}
E\{\Gamma_2\} = E\{\Upsilon_1\} + \frac{1}{\sum_{j=1}^{9}\pi_{F_j}H_{F_j}}\Bigg[\frac{1}{2}\Big[\Big(T_{\beta 2}-T_{\text{on}2}\Big)\pi_{F_3}H_{F_3}+\pi_{F_4}H_{F_4}+\pi_{F_5}H_{F_5} \\
+\Big(T_{\gamma 1}-T_{\text{on}1}+1\Big)\pi_{F_7}H_{F_7}+\Big(T_{\beta 1}-T_{\text{on}1}+1\Big)\pi_{F_9}H_{F_9}\Big] \\
+\frac{1}{3}\Bigg[\frac{\Big(T_{\gamma 1}-T_{\text{on}1}\Big)^2+6\Big(T_{\gamma 1}-T_{\text{on}1}\Big)+11}{\Big(T_{\gamma 1}-T_{\text{on}1}\Big)+5}\pi_{F_6}H_{F_6} \\
+\frac{\Big(T_{\beta 1}-T_{\text{on}1}\Big)^2+6\Big(T_{\beta 1}-T_{\text{on}1}\Big)+11}{\Big(T_{\beta 1}-T_{\text{on}1}\Big)+5}\pi_{F_8}H_{F_8}\Bigg]\Bigg].
\end{aligned}
\tag{218}
$$

As mentioned previously, the first term represents the waiting time in a system with no DRX. Hence, the second term denotes the additional waiting time caused by the cross-carrier-aware DRX.

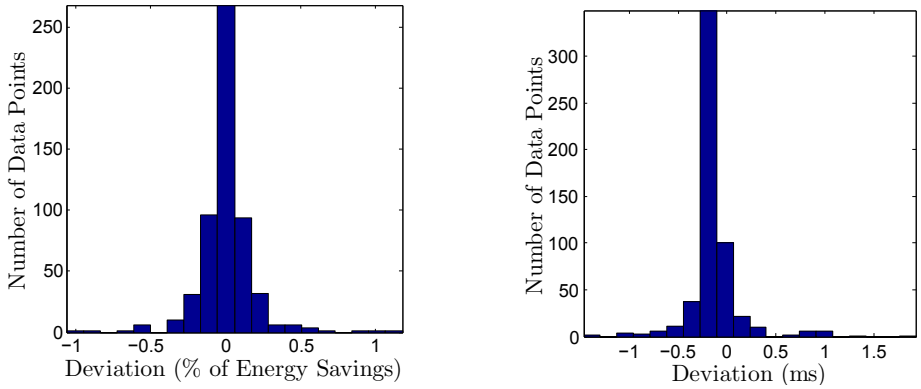
4.3.6 Performance Evaluation

The focus of this section is to show the validity and accuracy of our modeling approach, and to characterize the benefits provided in terms of energy savings and delay by our cross-carrier-aware DRX compared to the classical DRX.

To validate our modeling approach, we simulated a DRX system consisting of a BS and a UE. The link between them is composed of an anchor CC and a SCell operating according to our cross-carrier-aware DRX. The BS has an infinite buffer per CC per UE, where it stores any packet that cannot be immediately sent to the UE since the corresponding CC is in a "sleep" or "deep sleep" state, or because previously buffered packets are being sent. Once the CC is "awake," its buffered packets are sent by the BS following a FIFO scheme. We consider that there is no packet loss or retransmissions between the BS and the UE. The number of packets that arrive in a single subframe for the anchor CC and the SCell are denoted by Λ_1 and Λ_2, respectively, and their means are represented by λ_1 and λ_2. Similarly, the service times for the anchor CC and the SCell are denoted by X_1 and X_2, respectively, and their means are represented by b_1 and b_2. Λ_1 and Λ_2 are considered to

follow a Poisson distribution, and X_1 and X_2 are considered to follow a modified Poisson distribution whose PMF is described by Eq. (126).

In Figure 39, we depict the histogram of the deviation between the analytical and simulation results for the energy savings and delay metrics of the SCell, across multiple SCell DRX parameters. For each combination of those parameters, the operation of the cross-carrier-aware DRX during 1 million subframes was simulated. At the end of each simulation, the energy savings and delay metrics were computed and compared to the results found from the analytical expressions. For each configuration, the deviation is then computed as the difference between the theoretical and simulated performance metrics.



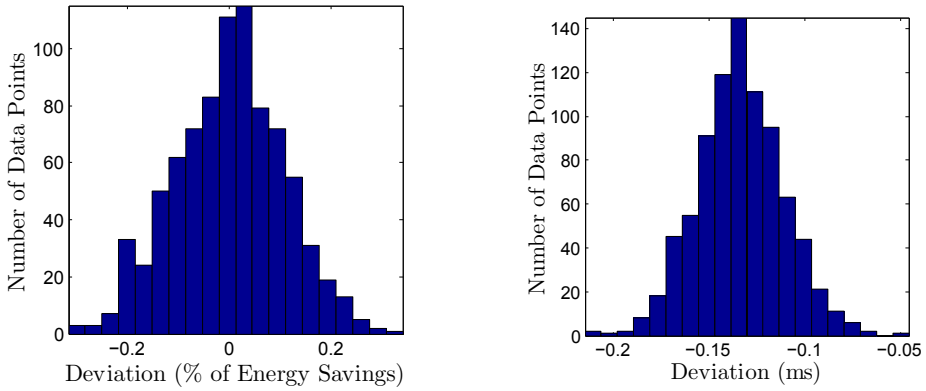
(a) Energy savings.

(b) Delay.

Figure 39: Deviation of theoretical from experimental metrics for the cross-carrier-aware DRX with parameters $\lambda_1 = 0.1$, $b_1 = 2.5$ms, $T_{\alpha 1} = 4$ms, $T_{\beta 1} = 8$ms, $T_{\text{on}1} = 2$ms, $T_{\gamma 1} = 16$ms, $N = 4$, $\lambda_2 = 0.1$, $b_2 = 2.5$ms, $T_{\alpha 2} \in [4, 8, 16, 32, 64]$ms, $T_{\beta 2} \in [4, 8, 16, 32, 64, 128, 256]$ms, $M \in [2, 4, 8, 16]$, $T_{\text{on}2} \in [2, 4, 8, 16, 32, 64, 128]$ms.

From Figure 39, we observe that the deviation for both metrics is extremely low. In particular, the absolute deviation in the energy savings is less than 1% of energy savings. Similarly, the absolute deviation in the delay is mostly within 1ms. From these results, we validate the significantly high accuracy, with respect to the DRX parameters of the SCell, of the analytical expressions derived for the performance metrics of the cross-carrier-aware DRX.

In Figure 40, we depict the histogram of the deviation between the analytical and simulation results for the energy savings and delay metrics of the SCell across multiple DRX parameters for the anchor CC. For each combination of those parameters, the deviation is computed as for Figure 39. Here, we also have extremely low deviations. In particular, the absolute deviation in the energy savings is less than 0.3% of energy savings. Similarly, the absolute deviation in the delay is less than 0.25ms. From these results, we validate the significantly high accuracy, also with respect to the DRX parameters of the anchor CC, of the analytical expressions derived for the performance metrics of the cross-carrier-aware DRX.

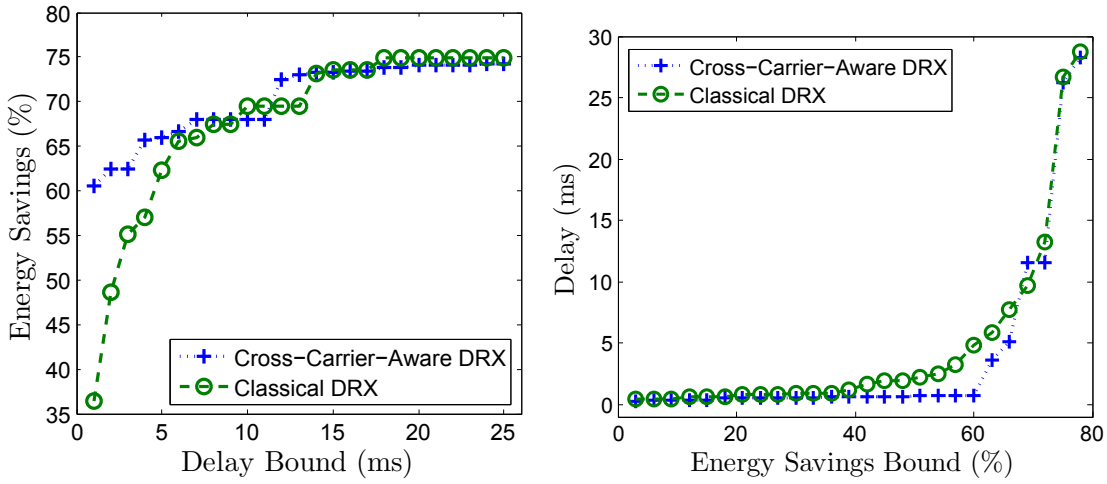


(a) Energy savings.

(b) Delay.

Figure 40: Deviation of theoretical from experimental metrics for the cross-carrier-aware DRX with parameters $\lambda_1 = 0.1$, $T_{\alpha 1} \in [4, 8, 16, 32, 64]$ms, $b_1 = 2.5$ms, $T_{\beta 1} \in [4, 8, 16, 3, 64, 128]$ms, $T_{\text{on}1} = [2, 4, 8, 16, 32, 64]$ms, $T_{\gamma 1} \in [1, 2]T_{\beta 1}$, $N \in [2, 4, 8, 16]$, $\lambda_2 = 0.1$, $b_2 = 2.5$ms, $T_{\alpha 2} = 4$ms, $T_{\beta 2} = 32$ms, $T_{\text{on}2} = 16$ms, and $M = 2$.

We now compare the improvements in the performance metrics of the SCell provided by our cross-carrier-aware DRX to those of the classical DRX across multiple DRX parameters by directly using the analytical expressions we have derived. For every possible combination of DRX parameters and a given maximum delay, we compute the highest energy savings provided by the classical and our cross-carrier-aware DRX. Similarly, for a minimum energy saving, we compute the minimum delay caused by both DRX schemes.



(a) Energy savings.

(b) Delay.

Figure 41: Difference in the performance metrics of the cross-carrier-aware DRX over the classical DRX. SCell parameters $\lambda_2 \in [0.05, 0.1]$, $b_2 = 2.5$ms, $T_{\alpha 2} \in [4, 8, 16, 32, 64]$ms, $T_{\beta 2} \in [4, 8, 16, 32, 64, 128, 256]$ms, $T_{\text{on}2} \in [2, 4, 8, 16, 32, 64, 128]$ms, $M \in [1, 2, 4, 8, 16]$, and for the classical DRX $T_{\gamma 2} = 2T_{\beta 2}$. Anchor CC parameters $\lambda_1 = 0.125$, $b_1 = 2.5$ms, $T_{\alpha 1} \in [4, 16, 64]$ms, $T_{\beta 1} \in [4, 32, 256]$ms, $T_{\text{on}1} \in [2, 16, 128]$ms, $N = 1$, $T_{\gamma 1} = T_{\beta 1}$.

Figure 41a depicts the energy savings provided by our DRX and the classical DRX, while Figure 41b does a similar comparison for the delay. In Figure 41a, we observe that our cross-carrier-aware DRX significantly outperforms the classical DRX when the delay limit is less than 5ms. For a higher delay limit, the difference between the two DRX schemes is very small. However, increasing the energy savings while maintaining a low delay limit is what presents the greatest challenge; thus, the cross-carrier-aware DRX proves to be much more efficient than the classical scheme.

In Figure 41b, we observe that our cross-carrier-aware DRX also outperforms the classical DRX when the energy savings limit is up to 60%. Over the interval of 60% to 80%, the delay caused by the cross-carrier-aware DRX is not significantly different from that of the classical DRX.

4.4 Conclusions

Because of its limited on-board energy, it is critical for the UE to maximize its energy efficiency. With this objective in mind, 3GPP introduced in LTE the use of DRX to minimize the energy consumption at the UE. For scenarios that support carrier aggregation in LTE-A, the use of DRX still remains the best approach to reducing the UE energy consumption. However, simply using the classical DRX scheme, as typically done in the literature, is inefficient. In this chapter, we first developed a semi-Markov Chain model to characterize the operation and performance metrics of the classical DRX. Through extensive simulations, the analytical expressions for the performance metrics were shown to be highly accurate, and the impact of the multiple DRX parameters on such metrics was analyzed. Second, we proposed a novel cross-carrier-aware DRX for scenarios that support carrier aggregation. We developed a semi-Markov Chain model and obtained the analytical expressions for the performance metrics for our proposed DRX scheme. The accuracy of those expressions was validated through extensive simulations. Then, we compared the performance of our cross-carrier-aware DRX against that of the classical DRX. We found that our DRX scheme significantly outperforms the classical DRX in terms of energy savings, especially in the most challenging condition of low tolerable delay. Moreover, the delay caused by our cross-carrier-aware DRX was not found to be significantly different from that of the classical DRX.

CHAPTER 5

REDUCING THE ENERGY CONSUMPTION THROUGH SMALL CELLS

As we found in Chapter 2, the use of small cells is one of the most effective methods to reduce the energy consumption in HetNets. In particular, femtocells are expected to have the biggest impact among all small cell types [7]. However, their effectiveness is affected by the cross- and co-tier interference, the backhaul bottleneck, service degradation due to user mobility, and privacy concerns. To address these issues, we developed a novel small cell solution, the femtorelay. Since the femtorelay needs to interoperate with the already existing elements in cellular networks, we developed the techniques and procedures to achieve its transparent integration with legacy cellular networks. To demonstrate the feasibility of such integration and the advantages of the femtorelay over regular femtocells, we developed models of the femtocell and the femtorelay in OPNET, a high-fidelity tool for modeling, simulation, and analysis of wireless networks, widely used in industry and academia. Performance evaluation is provided demonstrating the benefits of using femtorelays in terms of not only reduced energy consumption, but also decreased interference and increased capacity.

5.1 Motivation and Related Work

Co-channel deployments of macrocells and small cells are affected by cross- and co-tier interference. The cross-tier interference occurs because of macrocell signals, both uplink and downlink, interfering on the ones of nearby small cells, and vice versa. This interference is the main limiting factor in closed-access deployments [88]. Co-tier interference occurs between small cells in close proximity. Such interference can be a major issue in dense deployments, such as enterprise scenarios. In addition to the interference issues, the backhaul may severely affect the performance of small cells. Particularly, an internet-based

backhaul may suffer congestion, resulting in a backhaul bottleneck [89]. Under these conditions, the backhaul is unable to consistently provide the required QoS to satisfy the user expectations. The presence of small cells may also lead to service degradation or even interruption as mobile users experience frequent handovers from macrocells to small cells, from small cells to macrocells, and among small cells. Beyond the previously mentioned issues, privacy is also a major concern when privately owned small cells, such as femtocells, are open to macrocell users and, therefore, the femtocell owner's private network is exposed to traffic from unknown and potentially malicious users.

Although our femtorelay patented technology [90] [91] described in this chapter proposes a novel approach to address the aforementioned issues by exploiting femtocells and relays synergistic advantages, several other works in the literature have aimed to obtain some gain from the combination of femtocells and relays in cellular networks. In [92], [93], and [94], femtocell users (fUEs) act as relays for macrocell users (mUEs) by relaying the data of the latter to the serving femtocell. In [95], the authors proposed a concept where signal decoding is performed not at the femtocell, but at the macrocell base station (MBS), and both fUEs and mUEs can utilize the open-access single-backhaul femtocell. Works targeted to improve backhaul performance by using both relay and femtocell internet-based backhauls have also been proposed. In [96], the femtocell backhaul is used to aid the macrocell backhaul when the latter is congested, and [97] proposes a rate-splitting approach where mUEs can share their load between the direct macrocell link and an open-access femtocell using an internet-based backhaul. Femtocells and relays can also coordinate to improve mobility or resource allocation performance, as shown in [98] and [99], respectively. All these concepts are very different from ours since no new small cell technology is proposed in any of them. Furthermore, these concepts suffer from the need of re-architecting the network. Only the authors of [100] use a similar concept to the one we propose for a femtocell and relay cooperation framework, but they assume LTE relay-enabled networks; hence, their concept is not applicable to 3G networks. Further, femtocells need to be underutilized

to be useful for mUEs, since no additional mechanism for optimizing femtocell resources is applied. Moreover, no self-interference mechanism is applied for an optimal usage of the available resources.

The rest of this chapter is organized as follows. In Section 5.2, we describe our femtocell development platform created in OPNET to quantify the effects of interference on the performance of femtocells. In particular, in Section 5.2.1, we describe the new elements introduced by 3GPP in the cellular network architecture to support femtocells. In Section 5.2.2, we detail the modeling work performed in OPNET to enable the support of femtocells. We present the results of evaluating the performance of the femtocell-enhanced network in Section 5.2.3. In Section 5.3, we present our novel small cell solution, the femtorelay. In particular, the concept and functional descriptions are presented in Sections 5.3.1 and 5.3.2, respectively. Utilizing the models described in Section 5.3.4, we evaluate, in Section 5.3.5, the performance of the femtorelay against that of femtocells, showing the potential of our solution to overcome the limiting factors of femtocells. In Section 5.3.6, we introduce the concept of Multi-Femtorelay as an evolution of our technology intended for large-scale indoor environments. Concluding remarks are presented in Section 5.4.

5.2 Femtocell Development Platform

The detailed modeling and corresponding implementation of this platform is where the value of the work described in this section resides. Moreover, this platform establishes the grounds for validating the feasibility and performance of our femtorelay. To the best of our knowledge, no research tool has been developed to such level of detail, capturing both the physical and architectural effects of femtocell networks. In addition, communication between entities always takes place according to UMTS legacy and new femtocell-supporting protocols.

5.2.1 Femtocell-Enabled Architecture

In addition to the typical elements found in the packet switched (PS) section of a UMTS network, namely Node-B, radio network controller (RNC), serving GPRS support node (SGSN), and gateway GPRS support node (GGSN), two new main entities must be added to support femtocells [101] [102]: the home Node-B (HNB) and the HNB gateway (HNB-GW), as depicted in Figure 42.

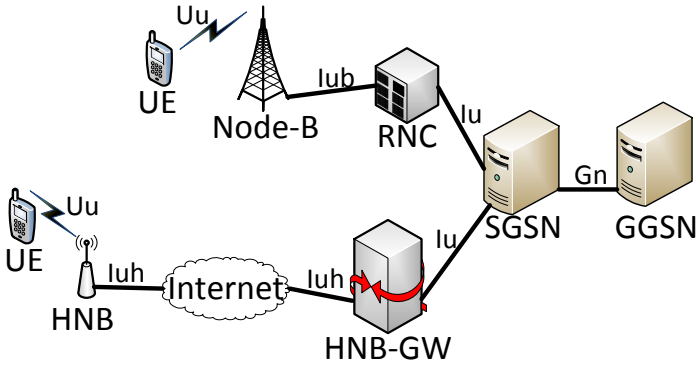


Figure 42: UMTS network - packet switched section.

The HNB-GW appears as a typical RNC towards the UMTS core network (CN) (i.e., SGSN and GGSN for the PS section). Therefore, to communicate with the CN, the HNB-GW utilizes the same Iu interface [103] as the one used by RNCs for this purpose (Iu-PS towards the PS section and Iu-CS towards the circuit switched section). The main role of the HNB-GW is to act as an aggregator for all the HNB connections. On the other hand, the HNB appears as any other Node-B to the UEs, providing the same Uu air interface as Node-Bs do. Between the HNB-GW and the HNB, a new interface, the Iuh [104], is defined to support HNB-specific procedures and a lightweight mechanism for carrying data and control traffic between the HNB-GW and the HNBs.

The new Iuh interface can be seen as a simplified version of the Iu interface, as shown in Figure 43. It removes the heavy-weight protocols from the Iu interface and introduces two new protocols: radio access network application part -RANAP- user adaptation (RUA) signaling [105], and home Node-B application part (HNBAP) [106]. The main objective

of RUA is to transparently transfer RANAP messages between the CN and the HNB, using the HNB-GW as intermediate entity. On the other hand, the main objective of HNBAP is to support the (de-)registration of HNBs and UEs and relocation of radio network subsystem application part (RNSAP). Both protocols utilize specific procedures to support these functionalities, and their respective messages are defined by 3GPP through abstract syntax notation one (ASN.1) [107] [108].

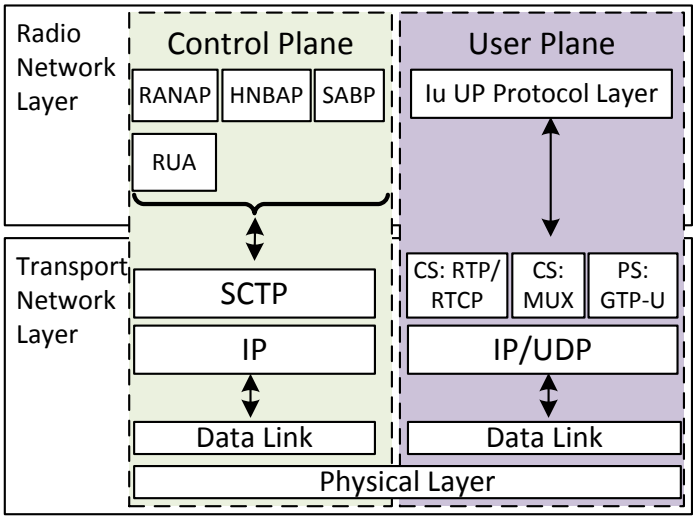


Figure 43: Iuh interface.

5.2.2 Femtocell Model and Implementation

In OPNET, implementing the HNB and HNB-GW involved (i) development at the node, process, pipeline, and message levels, and (ii) adapting and integrating the existing elements of UMTS in OPNET.

5.2.2.1 Node Level

At the node level, we defined the main components that would integrate the HNB and HNB-GW, as well as how those components are interconnected through packet streams. As shown in Figure 44, the HNB node model is composed of a radio transmitter and a receiver to communicate with the UEs, an *hnb* process module containing most of its functionality (including the HNBAP protocol), and a *gtp* process module for the Iuh interface towards the HNB-GW. Two additional processes are integrated: one in charge of executing the RUA

protocol, and another as a multiplexer enabling future sectorized femtocells. In addition, the HNB node model contains a standard IP protocol stack (not shown in the figure).

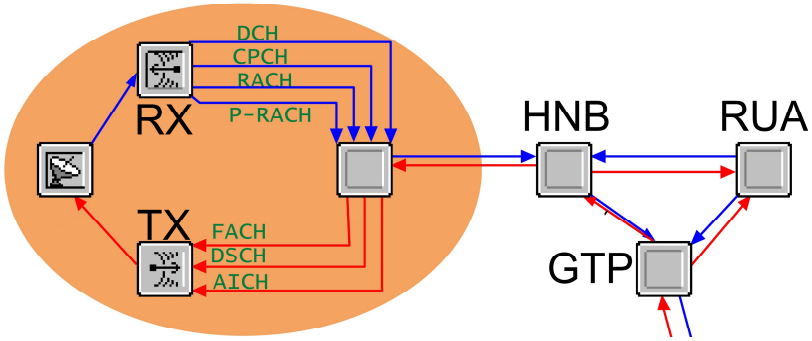


Figure 44: Section of the HNB node model.

At the HNB-GW, the focus of the implementation was to support the termination of the RUA and HNBAP protocols and the management capability for multiple femtocells. With the exception of RUA, all the above was integrated into an *hnb-gw* process module, as shown in Figure 45. The IP protocol stack (not shown in the figure) is also present, as well as the *gtp* process module. Here, the *gtp* module takes care not only of the Iuh interface towards the HNB, but also of the Iu-PS interface towards the core network.

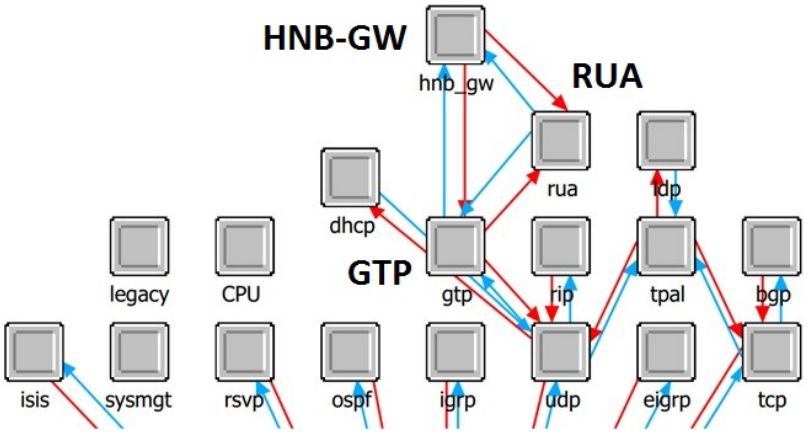


Figure 45: Section of the HNB-GW node model.

5.2.2.2 *Process Level*

The operation of each process module is defined at the process level as a state transition diagram, as the one shown in Figure 46. Internally, each diagram has the C code that determines the conditions for each transition, the procedures executed before a transition takes place, and the procedures executed within each state. At this level, the core functionality of the models is implemented. For the *hnb* process module, this functionality includes most of the Node-B and RNC services, including the mobility and resource management. At the *hnb-gw* process module, the implementation at this level focuses on adding the capability of supporting multiple femtocells and the HNBAP. For the latter, special attention is given to the support of the UE registration procedure, which enables the registration of a UE at the CN through the HNB and HNB-GW. For the *rua* process module, we implemented the direct transfer and HNB originated connect procedures, as defined in [105]. These two procedures support the establishment of an initial connection for the UEs and the subsequent exchange of RANAP control signaling between the HNB and HNB-GW.

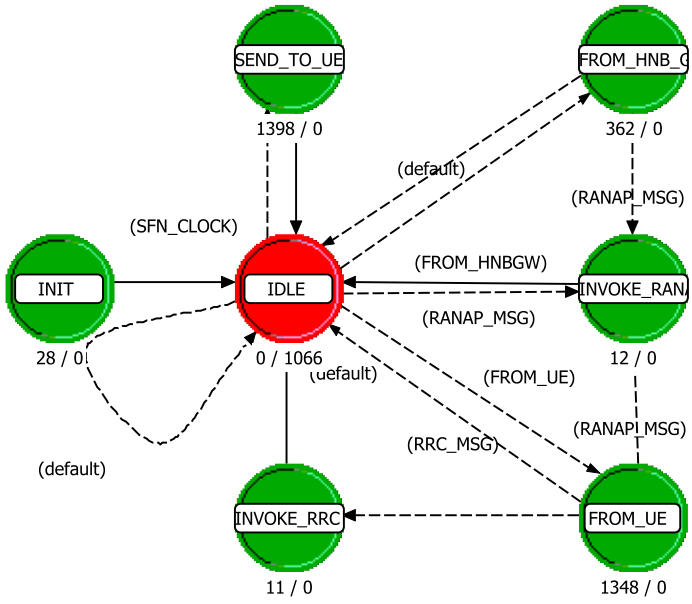


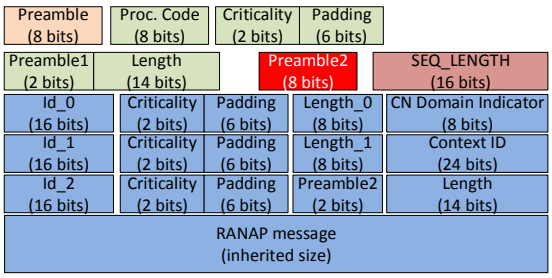
Figure 46: HNB process - state diagram.

5.2.2.3 *Message Level*

Each of the procedures implemented for the RUA and the HNAP utilizes unique message formats defined by 3GPP utilizing ASN.1. ASN.1 provides an abstract syntax of the structure and format of the information elements of the message, as shown in Figure 47a. 3GPP specifies the use of ASN.1 basic packed encoding rules (BASIC-PER) aligned variant as transfer syntax of its messages. By utilizing the ASN.1 message definitions of the RUA and HNBAP procedures and applying the PER encoding, we created the equivalent message format in the packet notation of OPNET, as shown in Figure 47b. The messages for which this conversion was applied include HNBAP UE register request, HNBAP UE register accept, HNBAP UE register reject, HNBAP UE deregister, UMTS RUA connect, and UMTS RUA direct transfer.

```
LocUpdatingAccept ::= SEQUENCE {
   locAreaIdent LocAreaIdent,
   --<iei format="tlv">17</iei>
   mobileIdentity MobileIdentity OPTIONAL,
   /* followOnProceed and ctsPermission contain
      present flags are sufficient to indicate
   --<iei format="t">A1</iei>
   followOnProceed NULL OPTIONAL,
   --<iei format="t">A2</iei>
   ctsPermission NULL OPTIONAL,
   --<iei format="tlv">4A</iei>
   equivPLMNs PLMNList OPTIONAL,
   --<iei format="tlv">34</iei>
   emergNumList EmergencyNumberList OPTIONAL
```

(a) Message specification in ASN.1.



(b) Message specification in OPNET's packet format.

Figure 47: HNB message specification.

5.2.2.4 *Pipeline Level*

In OPNET, fourteen pipeline stages at the radio transceiver are used to capture the physical layer effects of the wireless transmissions. The pipeline stages were upgraded so that the femtocell and the macrocell accounted, at the performance metrics calculations, for the effects of the co- and cross-tier interference. Therefore, our platform enables the development, implementation, and testing of interference mitigation techniques. Furthermore, it

supports the three access modes defined by 3GPP:

- Open mode: Any user can connect to the femtocell.
- Closed mode: Only a specific set of users, the ones belonging to the closed subscriber group (CSG), is allowed to connect to the femtocell.
- Hybrid mode: Any user can connect. However, users belonging to the CSG receive prioritized service.

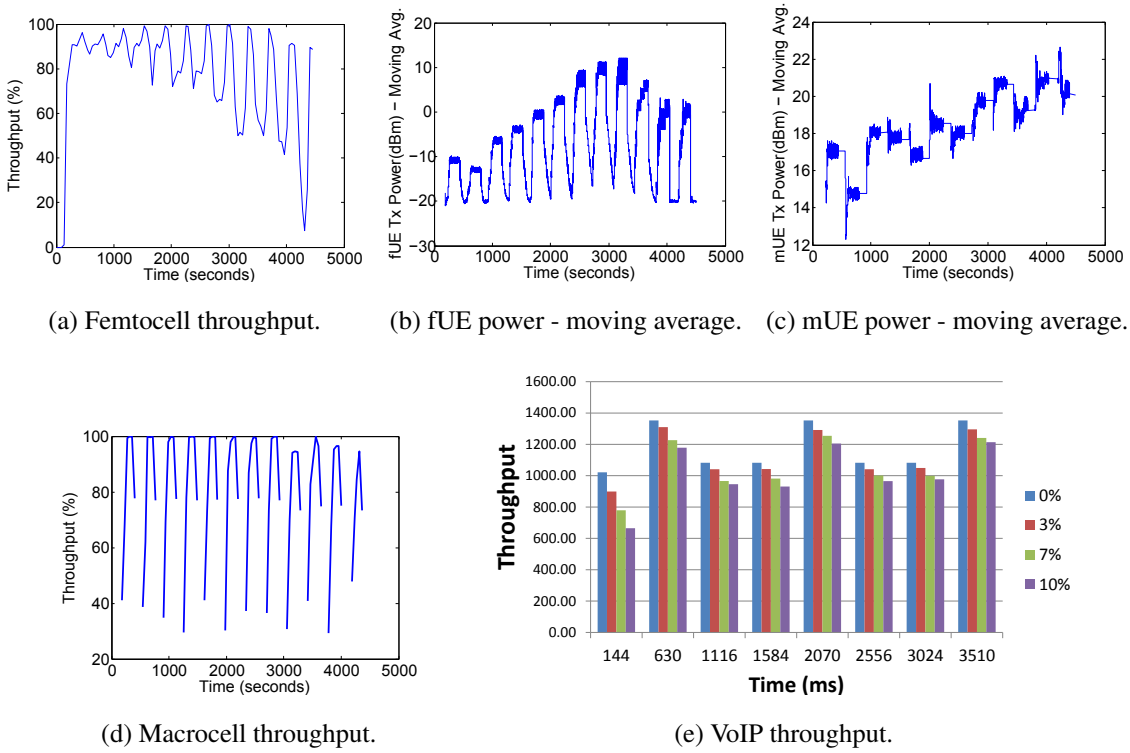
5.2.2.5 Integration with Exiting Components

To fully integrate the femtocell elements with the existing UMTS models in OPNET, we modified several of those models. At the UE, we added the support of CSG and of cell information acquisition from the HNBs. At the SGSN, we added procedures to discover the HNB-GWs and perform CSG-based admission control. At the GTP protocol, we enabled the automatic and transparent establishment of the tunnels with the existing nodes of UMTS.

5.2.3 Performance Evaluation

Since the performance metrics of a femtocell are mostly limited by the cross-tier interference and the backhaul reliability, our platform was designed to support the configurability of these parameters, as shown in Figure 48. Particularly, Figure 48a depicts how the throughput at a closed-access femtocell reduces as a macrocell user (mUE) approaches the femtocell, while intermittently transmitting to its own macrocell BS. This behavior occurs because the interference increases as the path loss between the mUE and the HNB decreases. The femtocell user (fUE) compensates for the higher interference by increasing its own transmission (TX) power, as shown in Figure 48b. However, since its maximum TX power is capped, the fUE cannot compensate for all the interference; therefore, its throughput reduces. As a result, the fUE does not meet its QoS requirements, even by maxing out its energy consumption. On the other hand, the mUE is also affected by the interference

from the fUE. Particularly, Figure 48c shows the transmission power of the mUE increasing to compensate for the interference caused by the fUE. By the end of the simulation cycle, the mUE reaches its maximum transmission power and can no longer compensate for the fUE interference; therefore, its throughput starts to decrease, as shown in Figure 48d.



(a) Femtocell throughput. (b) fUE power - moving average. (c) mUE power - moving average.

(d) Macrocell throughput. (e) VoIP throughput.

Figure 48: Femtocell performance metrics.

For a different simulation scenario, Figure 48e shows how packet losses of 3%, 7%, and 10% at the backhaul bring the average bit rate of a VoIP call to 96%, 91% and 88% of the baseline, respectively. This loss rate is higher than the 1% expected for VoIP [109].

5.3 Femtorelay

The femtorelay (FR) is our novel and cost-effective small cell technology conceived to improve the capacity, coverage, backhaul resiliency, and energy efficiency in indoor environments beyond that of existing solutions [91]. The patented femtorelay system architecture [90] provides the ground of our innovation in the areas of relaying, backhaul management, cross-tier and self-interference cancellation, and co-tier deployments optimization.

5.3.1 Concept Description

The supporting architecture for the femtorelay is depicted in Figure 49. The femtorelay access point (FrAP) is the attachment point for nearby users. It has dual-backhaul connection to the CN: an internet-based and a relay-based one. The internet-based is supported at the RAN by the HNB-GW, which aggregates the traffic of multiple FrAPs. The relay-based utilizes the same RF frequency as the one used by the BS to communicate with the mUEs, i.e., it acts as an in-band relay link [110] [111] [112] and is supported at the CN by the femtorelay gateway (FrGW). The role of the FrGW is similar to that of the HNB-GW, i.e., to transparently support the integration of multiple FrAP into the network. To serve the users, a hybrid-access mode is utilized, where a set of fUEs is pre-authorized to access the FrAP and all its resources, and a set of mUEs is allowed to attach, but with certain restrictions on the resources those UEs are permitted to use [113] [114] [115].

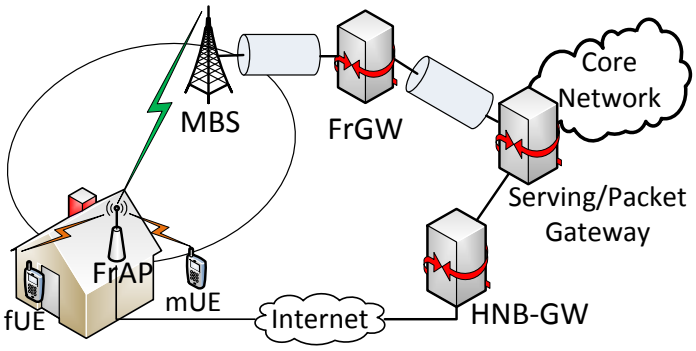


Figure 49: Femtorelay network architecture.

Internally, the FrAP consists of the major blocks shown in Figure 50. The femto module

manages the radio link towards the mUEs and fUEs; the relaying module manages the wireless backhaul towards the macrocell BS; the smart resource management allocates the resources to both fUEs and mUEs along the access link; the performance monitor tracks the user throughput across the backhaul connections; the self-interference cancellation enables full duplexing at the backhaul and access links; the mobility management enables handover and backhaul switching for the users.

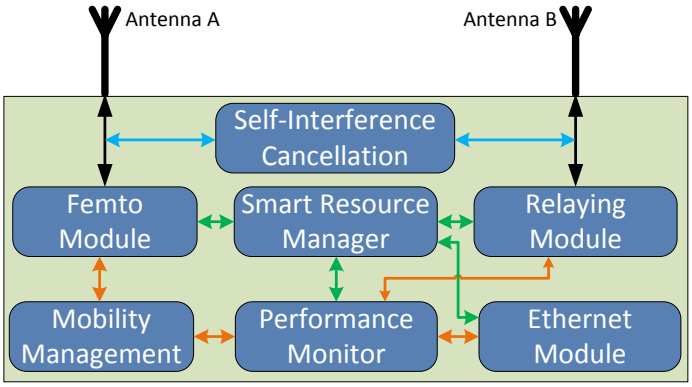


Figure 50: Femtorelay concept description.

5.3.2 Functional Description

There are significant advantages of deploying femtorelays, as it will be further elaborated in the rest of this section. The crucial problem of interference between macrocell and small cells is overcome by serving the interfering mUEs residing within the small cell coverage area directly from the small cell through the relay-based backhaul. The dual-backhaul feature also promises improved QoS for users and supports novel functions including backhaul switching to enhance user experience. The flexible femtorelay approach of utilizing multiple backhauls to access the core network and minimize the interference improves the QoS of the users and is a unique feature of our technology.

Scalability and Compatibility: To facilitate a rapid large-scale adoption of our technology, we introduce the FrGW, as shown in Figure 49. Its main role is to significantly minimize the impact at the RAN, UE, and CN, given the presence of the FrAP relay module. As part of the operator's CN, the FrGW provides standard interfaces towards the rest

of the elements of the network, transparently supports and aggregates the traffic from multiple FrAPs, and encapsulates and decapsulates control and data messages for the FrAPs and their attached users. Moreover, the FrGW enables the standard-compliant support of not only femtorelays in UMTS and LTE networks, but also relays, which would otherwise not be possible in UMTS and would require extensive upgrades in LTE.

Backhaul Link Establishment: Using our novel backhaul establishment procedure and the FrGW, the relay backhaul from the FrAP to the neighboring macrocell BS is established as follows. The relay module of the FrAP acts as a mobile user and establishes a radio bearer with the macrocell. Such bearer is utilized to then establish a default UMTS radio access bearer (RAB) towards the FrGW, i.e., the Fr-GW is executing the role of the SGSN for the mobile user identity of the FrAP. The FrAP can utilize this default session to exchange the backhaul signaling messages for user attachment.

Session Establishment Procedure: The session establishment procedure can be explained as follows. The FrAP transmits downlink synchronization signals that are received by users who can then perform initial cell attachment to establish the radio access link. With the help of user classification and the novel resource manager, a decision can be made regarding whether the users are served through the internet-based backhaul or the relay backhaul. If the internet-based backhaul is selected, the session is established in the same way as would be done in a femtocell. In the case where the users are served by the relay backhaul, the default session setup from the relaying module with the network is utilized to negotiate the session with the core network for the user. The session establishment procedure is completed when the end-to-end connectivity is achieved for the mobile users with the core network.

Interference and Backhaul Management: As shown in Section 5.2.3, an mUE causes interference to nearby closed-access femtocells and vice versa, which leads to a service degradation of both. On the other hand, the FrAP serves the strongly interfering mUEs and internally processes their traffic towards its relaying module. The relaying module

forwards the traffic through the relay-based backhaul to the CN. By managing the traffic of the mUEs, the FrAP achieves two objectives: it (i) significantly reduces the cross-tier interference between the macrocell and the FrAP, and (ii) avoids the congestion caused by the mUE traffic at the private internet backhaul.

QoS Guarantee and Mobility: In femtocells, congestion at the backhaul significantly degrades the QoS provided to the users, as we showed in Section 5.2.3. This problem is largely mitigated at the FrAP by the interactions between the performance monitor and the mobility management unit and their handling of the FrAP backhauls. When the performance monitor detects that the backhaul currently used by a UE is not capable of satisfying the QoS requirements (typically the internet-backhaul would be the one in such situation), it determines if a different backhaul can satisfy such requirements or if the UE must be handed over to a nearby cell. This decision is passed to the mobility management unit that seamlessly executes the action indicated by the performance monitor: a dynamic backhaul switching or an outbound handover [116] [117] [118]. The mobility management unit also plays a key role in handover procedures from nearby cells to the FrAP.

Self-interference Cancellation: The relaying operation of the FrAP requires radio resources for communication with the macrocell. The macrocell allocates such resources to the relaying unit. The smart resource manager takes into account this information to assign radio resources to the users it serves at the access link. In addition, to achieve better spectral efficiency, each FrAP transmitter may operate in the same frequency band as its receiver and perform simultaneous transmission and reception for the access and relay links [119] [120] [121] [122] [123] [124], resulting in a full-duplex operation. Even though full-duplex communication can add capacity to wireless networks [125], it results in the phenomenon of self-interference, where a signal from the transmitter causes a strong interference at the receiver and makes signal extraction infeasible [126] [127] [128]. The FrAP includes a robust self-interference cancellation scheme to mitigate the undesired feedback signal and operate in full-duplex mode.

The self-interference feedback channel, as shown in Figure 51, exists between the femtorelay transmitter and receiver. For example, the BS will send downlink traffic to the FrAP on f_1 while the traffic is relayed to a UE on f_1. Similarly, the UE will send uplink traffic to the FrAP on f_2, which is simultaneously relayed to the BS on f_2. The FrAP receivers operating on f_1 and f_2 will suffer from self-interference from their respective transmitters. A similar problem occurs when the FrAP backhaul channel is the same as its access channel. The feedback channel can most often be modeled as an indoor line-of-sight (LOS) channel with additive white gaussian noise (AWGN). It has been shown that this channel varies slowly in time [129].

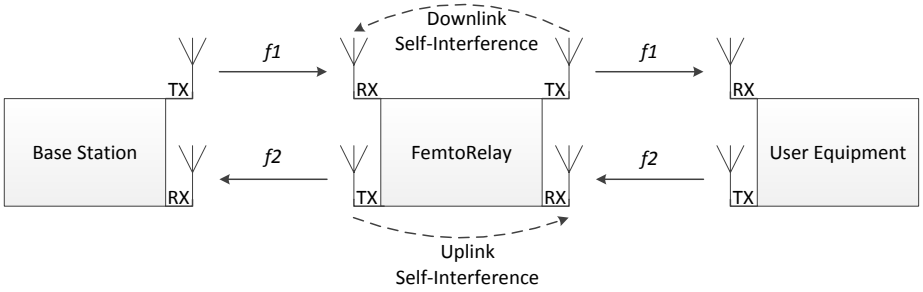


Figure 51: Self-interference channel in a femtorelay.

To achieve full-duplex capability, the FrAP includes a self-interference cancellation mechanism that includes a unique combination of active [130] [131] [132] [133], passive [134] [135], and physical self-interference cancellation techniques [136]; such combination blends the features of each method while mitigating their drawbacks to enable full-duplex operation across a wide bandwidth.

5.3.3 System Integration

As mentioned before, the FrAP has two backhaul connections. The internet-based backhaul is the same as the one utilized by a HNB. As such, it utilizes the same procedures, interfaces, and protocol stacks for the user and control plane as do femtocells [105][106] [137][138]. On the other hand, for the relay-based backhaul link, the protocols, procedures, and interfaces are different.

From the relay-based backhaul point of view, the FrAP and FrGW jointly provide an extra layer of tunneling that enables a transparent exchange of data and control messages between the UE and the CN, as shown in Figure 52.

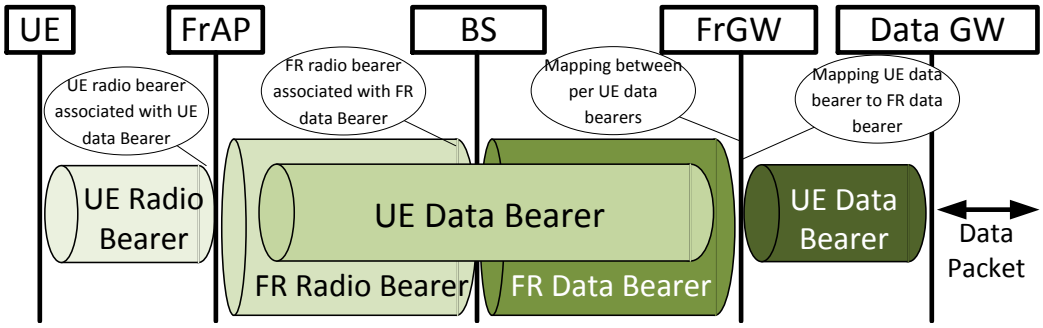


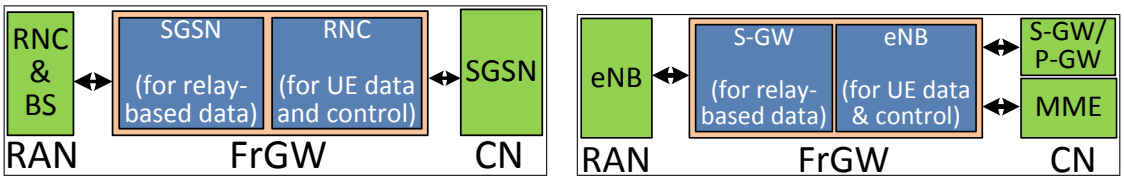
Figure 52: Tunneling for femtorelay.

The FrAP and FrGW extract and insert any message into this tunnel without loss of information. This is analogous to the virtual private networks (VPNs) established between two VPN endpoints. Via this tunnel, any protocol, procedure, and interface can be executed. For example, in the LTE version of the FrAP, this tunnel would support the X2 [137] and S1 [138] interfaces, also used by eNodeBs and home eNodeBs. In the UMTS/HSPA version, the tunnel would support the Iu [103] or Iuh [105][106] interfaces, also used by RNCs and HNBs, respectively. Thus, the FrGW plays a key role in supporting the femtorelay concept.

In general, adding a new element to the CN requires either standardizing this element or modifying the CN to make it compatible with the new element. Both options have their advantages and drawbacks. Without standardization of the new element, modifying the CN to be compatible with the new element represents a very high risk for an operator since there is no guarantee that the modifications will achieve the desired results without causing disruptions to the service. On the other hand, the standardization path takes a long time to achieve, but guarantees that all future networks will support the new element. Nevertheless, even if the standardization path took less time, operators are very cautious regarding upgrading their equipment to newer versions of a standard due to the high impact

of any system disruption. Therefore, the time taken by an operator to upgrade its equipment (after standardization is completed) would significantly delay the introduction of the new element. With this in mind, we chose an option that represents a balance between the previous two: comply with the existing standards by providing standard interfaces to the new elements we are introducing.

To avoid any modifications to the CN, the FrGW presents itself to the rest of CN elements as the RAN element that is serving the end user. In UMTS, this corresponds to the RNC and in LTE, to the eNodeB, as shown in Figure 53. This way, the rest of the CN elements can communicate with the FrGW through the interfaces and protocols they already use to communicate with RNCs and eNodeBs. Similarly, the FrGW presents itself to the RAN as the CN element that is serving the FrAP relay-based data. In UMTS, this corresponds to the SGSN and in LTE, to the S-GW. This way, the RAN can communicate with the FrGW using the same interfaces and protocols it already uses to communicate with SGSNs and S-GWs.



(a) FrGW compatibility with UMTS.

(b) FrGW compatibility with LTE.

Figure 53: FrGW compatibility with standards.

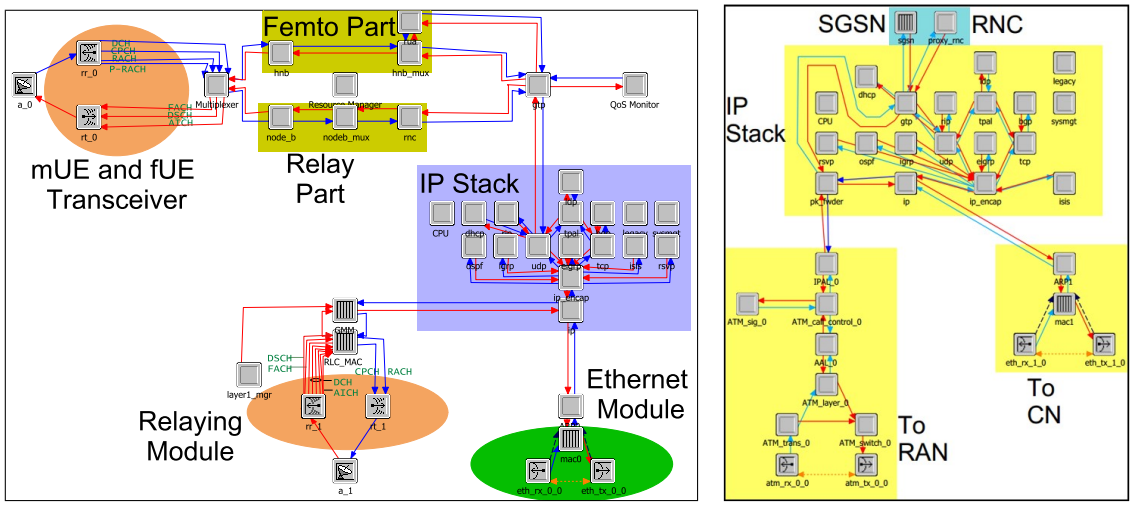
Even though we have described the FR architecture using 3GPP's terminology for UMTS and LTE, the FR is not restricted to these two specific radio access technologies (RATs). The previous descriptions can be mapped to other networks, such as WiMAX. Given the ability to perform such mapping, we do not go into more details regarding the specific protocols that will be used by the FR since they would depend on the RAT in which it is introduced.

5.3.4 Femtorelay Modeling

We implemented the FrAP and the FrGW in OPNET as a proof of concept to demonstrate that the FR can be integrated into a standard-compliant cellular network and to evaluate the performance of the femtorelay against that of closed-access femtocells. Similar to the femtocell OPNET work described in Section 5.2.2, designing a complete femtorelay model involved (i) development at the node, process, pipeline, and message level, and (ii) adaptation and integration of the existing elements of cellular networks in OPNET. Here, we focus on describing the node model of the FrAP and the FrGW.

The major sections of the FrAP node model are shown in Figure 54a. The mUE and fUE transceiver manages the physical layer aspects of the communication with users over the air interface of the access link. The femto part manages the traffic that is forwarded through the internet-based backhaul. The relay part manages the traffic that is forwarded through the relay-based backhaul. The IP stack includes all the IP protocols (e.g., TCP, UDP, DHCP). The relaying module manages the physical layer aspects of the communication with the BS over the relay-based backhaul. The ethernet module manages the communication over the internet-based backhaul. Except for the IP stack and ethernet module, we developed or customized each process module.

The major sections of the FrGW node model are shown in Figure 54b. As described in Section 5.3.3, the FrGW presents itself as an SGSN towards the RNC and as an RNC towards the SGSN. Therefore, such dual role is integrated as part of the FrGW OPNET node model. Both the SGSN and RNC process module of the FrGW are adapted to perform the encapsulation and decapsulation of the data and control traffic described in Section 5.3.3. As such, little-to-no modifications were introduced in the rest of the pre-existing models for cellular networks in OPNET.



(a) Femtorelay access point.

(b) Femtorelay gateway.

Figure 54: Femtorelay implementation in OPNET.

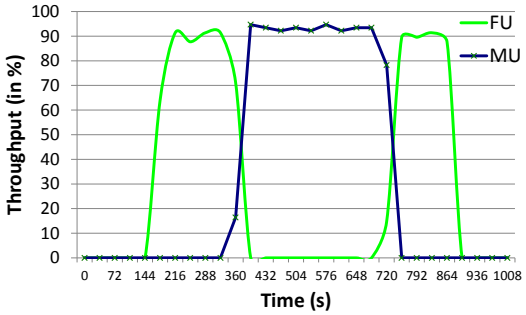
5.3.5 Performance Evaluation

Extensive simulations were performed to validate the performance of the femtorelay against that of closed-access (CSG) femtocells. Simulations were performed using different configurations by varying the transmit power of the access points (HNB and FrAP), user distances, and traffic types. In this section, we present our initial results where the throughput of mUEs and fUEs data transmission has been measured to see how it is affected when both users are in close proximity and transmit simultaneously under - 1) CSG femtocells, 2) femtorelay. The users locations are randomly distributed within the coverage area of the access points. The simulation parameters are listed in Table 9.

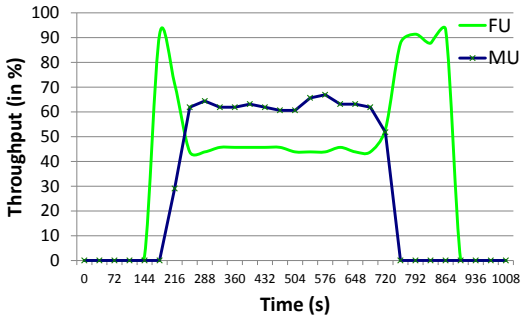
Table 9: Femtorelay simulation parameters.

User traffic type	*CBR*
User start times	$[100, 200]s$
User finish times	$[700, 1000]s$
BER required	10^{-3}
User traffic rate	100 kbps
Mean user distance	5 m

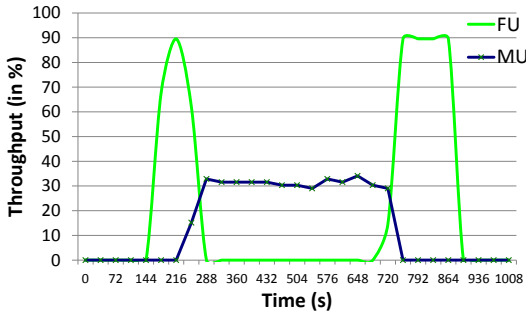
For a CSG femtocell scenario, Figure 55 depicts three major throughput outcomes that can occur when an ongoing fUE transmission experiences interference due to a transmitting mUE. The throughput performance of femtocells is computed as a percentage of the data transmission rate. In Figure 55a, the fUEs throughput drops to 0% as soon as the mUE starts transmitting, while the throughput of the mUEs remains as high as 90%. Once the mUE traffic comes to an end, the fUEs are able to achieve nearly 90% throughput. In Figure 55b, once the mUEs start transmitting, the throughput for both mUEs and fUEs is partially degraded. In Figure 55c, a different throughput performance is observed. Even though the throughput of the fUEs drops to 0% when the mUE starts transmitting, the throughput of the mUEs only degrades partially to 30%, compared to the 90% throughput achieved in Figure 55a. As seen from these results, the cross-tier interference has a significant role in the achievable performance in the network where CSG femtocells are present.



(a) fUE throughput degradation.



(b) fUE and mUE throughput degradation.

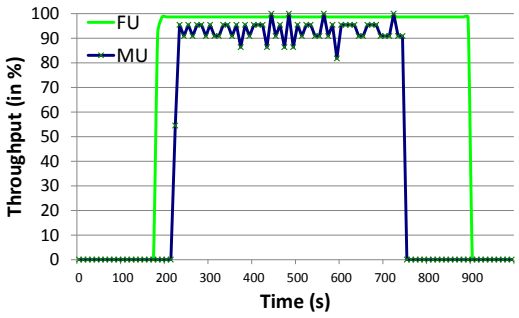


(c) Partial mUE and full fUE throughput degradation.

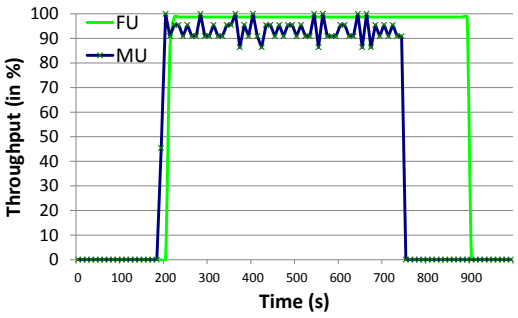
Figure 55: Throughput performance under femtocell.

For two scenarios where the distance from the FR to the mUE and fUE varies, Figure 56a and Figure 56b depict the throughput performance. In both cases, the fUE and the mUE achieve more than 90% of their throughput across the entire simulation duration, which was not possible in the femtocell case. Moreover, in contrast with the femtocell case discussed in Section 5.2.3, where the transmission power of the UE rapidly increases until reaching the maximum, Figure 56c shows that the average transmission power in the femtorelay scenario is fairly stable, only changing due to the fast power control implemented in UMTS. For the scenarios in Figure 56a and Figure 56b, we computed the mean total throughput, shown in Figure 56d. In the scenario where the femtocell would experience high interference due to a nearby mUE transmission, the FR provides around 3x higher mean total throughput. Even in the scenario where such interference was low, the FR provides around 1.5x higher mean total throughput. These results clearly indicate that the

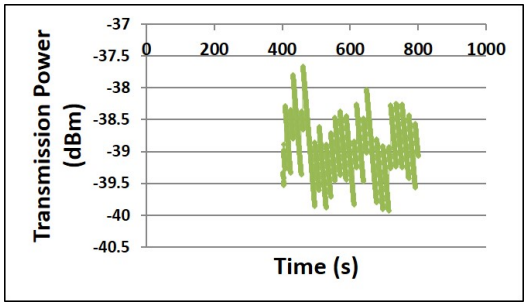
femtorelay is a strong candidate to overcome the cross-tier interference that inhibits the throughput performance in small cells, and particularly in femtocells. Such achievement increases not only the throughput performance, but also the energy efficiency provided by the use of small cells.



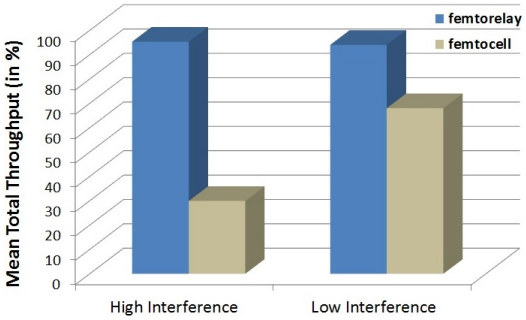
(a) Femtorelay - scenario 1.



(b) Femtorelay - scenario 2.



(c) Femtorelay transmission power.



(d) Mean total throughput - femtocell vs. femtorelay.

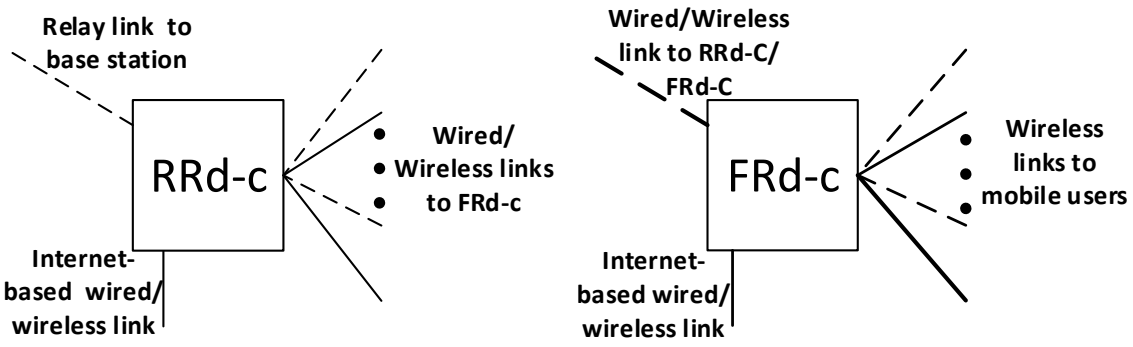
Figure 56: Femtorelay performance evaluation.

5.3.6 Technology Evolution for Large-Scale Indoor Environments

Femtocells were initially envisioned as devices installed by home users in their dwellings. As femtocells evolved, they were introduced in larger indoor environments, such as airports, office buildings, shopping malls, and stadiums to improve both capacity and coverage. Nevertheless, the fact that the femtocells are now deployed in a larger indoor environments does not eliminate the interference and congestion problems associated with them. With this in mind, we further expanded the femtorelay concept to address these two problems in larger indoor environments (enterprise environments from now on).

The Multi-Femtorelay (MFR) is the femtorelay version for enterprise environments [90]. The idea behind it is to have a central entity in charge of managing and coordinating the femtocells belonging to the enterprise environment, i.e., the ones being closely located. This approach provides better resource management across the femtocells compared to uncoordinated and distributed management approaches.

While the concept of having a central entity managing multiple femtocells can be mapped to the femtocell gateway (HNB-GW or HeNB-GW in 3GPP terms), the MFR differs in three things. First, the femtocell gateway is meant to be part of the CN of the service provider, while the central entity in the MFR is meant to be located within the local premises of the enterprise environment. This leads to the second difference: in the MFR, the central entity is in charge of managing a small set of femtocells (i.e., just the ones of the enterprise environment), compared to the number of femtocells that are managed by a femtocell gateway. This allows the MFR to optimize the performance of the enterprise femtocell network in ways that are computationally prohibitive for the HNB-GW to apply to all the femtocells in the network. Third, and what makes it unique, is that the central entity in the MFR has the same type of wired and wireless backhaul as the femtorelay, and, thus, brings all the benefits of the femtorelay to enterprise environments.



(a) RRd-C. (b) FRd-C.

Figure 57: Multi-Femtorelay components.

Figure 57 illustrates the main components of a MFR. First, there is the relay radio component (RRd-C). The main tasks associated with the RRd-C are to coordinate the usage of resources among the FRd-C and to provide both wireless and wired backhaul to the CN using similar methods as the femtorelay. The second element is the femto radio component (FRd-C). The main purpose of the FRd-C is to provide radio access to the mUEs and fUEs within the enterprise environment. The link between the RRd-C and the FRd-C can be established through a wired or wireless connection, depending on the scenario. For example, in an office building, the existing LAN can be used to connect the FRd-C to the RRd-C, as shown in Figure 58. In tunnels and stadiums, wireless links may be more appropriate if there is no pre-existing wired networks. Regarding the internet-based backhaul, the RRd-C, the FRd-C, or both may have it. For example, if all the FRd-C belong to the same owner, having the internet-based backhaul in the RRd-C is sufficient. However, if each FRd-C belongs to a different owner, then each FRd-C may have its own internet-based backhaul.

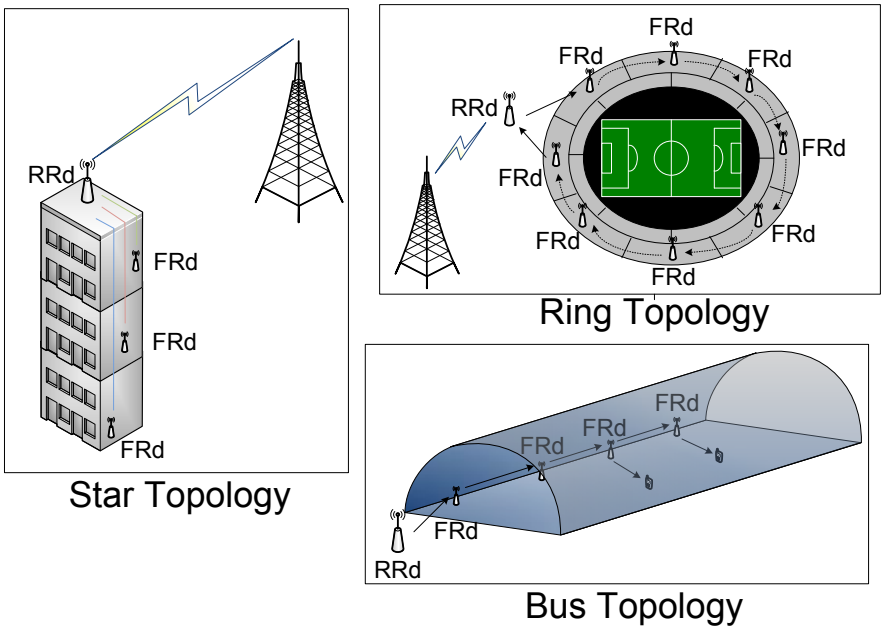


Figure 58: Multi-Femtorelay scenarios.

From the point of view of procedures, interfaces, and protocols, the MFR borrows these directly from the femtorelay. Therefore, the MFR is able to integrate seamlessly into the

network. In terms of managing the radio and backhaul resources, both the RRd-C and FRd-C have resource managers that closely interact and coordinate the resources that should be used by each element of the MFR. This coordination allows the MFR to establish efficient configurations within the enterprise environment to achieve multiple objectives, such as capacity optimization, mobility robustness, load balancing, and energy savings.

5.4 Conclusions

Even though the use of small cells is one of the most efficient techniques to minimize the energy consumption in HetNets, the effectiveness of small cells is severely hindered by the cross- and co-tier interference and by the backhaul unreliability. To characterize these effects, we created a complete femtocell modeling and development platform in OPNET that allowed us to quantify not only the throughput degradation, but also the increment in the energy consumption resulting from the interference between macrocells and small cells. To address these issues, we introduced a novel solution for next generation small cells, the femtorelay. We detailed not only the conceptual and functional descriptions of the femtorelay, but also its internal components, how it is able to integrate in a standard-compliant way with existing cellular systems, and how each of its benefits is achieved. In addition, we developed the models required to validate the feasibility of our technology and its performance against that of femtocells, depicting the potential of our technology to outperform existing solutions. Finally, a vision of the further evolution of this technology has been described, as well as its application to enterprise environments.

CHAPTER 6

CONCLUSIONS

Reducing the energy consumption in current and future heterogeneous cellular systems is of paramount importance not only from an environmental perspective, but also from an economic one. The traditional approach to addressing this problem has been to improve the energy efficiency of the hardware components. Nevertheless, such approach is not sufficient to sustain the exponential traffic growth that networks are experiencing. As a result, the evolution of cellular network technologies must shift from its traditional focus on coverage and capacity to one where the energy efficiency has a top priority.

In this thesis, we aimed at exploiting existing technologies and developing new ones to achieve energy savings not only for the operators, but also for the end users. In Chapter 2, we addressed the energy consumption at the RAN, where most of the energy is wasted. In light of the new technologies introduced in LTE-Advanced, in Chapter 3, we explored how multi-stream carrier aggregation can be utilized to further reduce the network energy consumption. In Chapter 4, we also leveraged such technologies to develop a new scheme of discontinuous reception to reduce the energy consumption at the user equipment. Finally, given the relevance of small cells in improving the network energy efficiency, in Chapter 5, we developed a novel small cell technology capable of providing not only higher capacity, but also reduced interference and energy consumption in HetNets. Now, we provide a summary of the contributions of each chapter.

In Chapter 2, we analyzed the energy consumption in multi-layer HetNets, accounting for its dependence on the spatio-temporal traffic variations of the traffic demands and on the internal hardware components. Then, we characterized finding an energy-efficient on-off and cell-association policy in terms of a Knapsack-like problem. For such problem, algorithms were developed to find a solution for the two- and m-layer cases. Results showed

that our algorithms consistently provided large energy savings across a wide range of scenarios and that small cells played a key role in achieving such savings during high traffic loads while large cells did so during periods of low traffic load.

In Chapter 3, we studied the problem of minimizing the energy consumption in MSCA-enabled HetNets and developed an efficient algorithm to solve it. We showed that, by utilizing a quasi-convex relaxation, we are able to not only solve the problem, but also to establish a clear and simple cell-association policy. Moreover, we demonstrated how this cell-association policy can be easily adjusted to obtain a new policy that balances the conflicting objectives of energy minimization and capacity maximization. Applying our algorithm, we obtained the energy-capacity trade-off curve and found that a large amount of energy savings can be achieved in an MSCA-enabled HetNet by slightly reducing the network capacity usage.

In Chapter 4, we developed a semi-Markov Chain model to characterize the operation and performance metrics of the classical DRX. Then, we proposed a novel cross-carrier-aware DRX for scenarios that support carrier aggregation. To analyze its performance, we developed a semi-Markov Chain model and characterized the performance metrics for our proposed DRX scheme. Then, comparing the performance of our cross-carrier-aware DRX against that of the classical DRX, we found that our DRX scheme significantly outperforms the classical DRX in terms of energy savings, especially in the most challenging condition of low tolerable delay. Moreover, the delay caused by our cross-carrier-aware DRX was not found to be significantly different from that of the classical DRX.

In Chapter 5, we introduced the femtorelay, our novel solution for next-generation small cells. We detailed not only the conceptual and functional descriptions of the femtorelay, but also its internal components, how it is able to integrate in a standard-compliant way with existing cellular systems, and how each of its benefits is achieved. In addition, we developed the models required to validate the feasibility of our technology and its performance against that of femtocells, depicting the potential of our technology to outperform

existing solutions.

In the future, our research in energy-efficient heterogeneous cellular systems will encompass the key technologies that are being introduced in next-generation cellular systems. One such technology is the use of higher frequencies in the millimeter wave (mm-Wave) band together with multi-stream carrier aggregation. In such scenario, the macrocell will behave as a mobility anchor operating over the traditional spectrum bands, while the small cells operate in the mm-Wave band. Therefore, the small cells can potentially be activated and deactivated on-demand as users request and release connections to the network, i.e., the number of cells in the network would dynamically adapt over very short time intervals to the existing users and their traffic. Even though such high-frequency adaptation would reduce the network energy consumption, it will open new challenge in terms of coordination methods, mobility estimation, resource management, and base station selection. Another technology that will further enhance the energy efficiency of HetNets is the use of massive MIMO. In such scenario, a large number of antennas is utilized to generate narrow beams toward every user. Nevertheless, existing work on the subject has mainly focused on exploiting massive MIMO to maximize the network capacity and not given enough attention to the impact on the energy consumption.

PUBLICATIONS

Journal

1. Chavarria Reyes, E., Akyildiz, I. F., "Cross-carrier-aware DRX for LTE-Advanced," submitted for journal publication, 2014.

2. Chavarria Reyes, E., Akyildiz, I. F., "Energy-efficient Multi-Stream Carrier Aggregation in HetNets," submitted for journal publication, 2014.

3. Chavarria Reyes, E., Akyildiz, I. F., and Fadel, E., "Energy Minimization Framework for Heterogeneous Wireless Systems," submitted for journal publication, Feb. 2014, revised Nov. 2014.

4. Akyildiz, I. F., Gutierrez-Estevez, D. M., Balakrishnan, R., Chavarria Reyes, E., Krier, J. R., "LTE-Advanced and the Evolution to Beyond 4G (B4G) Systems," Physical Communication (Elsevier) Journal, vol. 10, no. , pp. 31-60, March 2014.

5. Akyildiz, I. F., Chavarria Reyes, E., Gutierrez-Estevez, D. M., Balakrishnan, R., Krier, J. R., "Enabling Next Generation Small Cells through Femtorelays," in Physical Communications (Elsevier) Journal, Vol. 9, pp. 1-15, Dec. 2013.

6. Akyildiz, I. F., Gutierrez-Estevez, D. M. and Chavarria Reyes, E. "The Evolution to 4G Cellular Systems: LTE-Advanced," in Physical Communications (Elsevier) Journal, Vol. 3, No. 4, Dec. 2010.

Conference

1. Chavarria-Reyes, E. and Akyildiz, I.F., "Radio Access Network Energy Minimization in Heterogeneous Wireless Systems," in Proc. of the IEEE Personal, Indoor and Mobile Radio Communications (PIMRC), London, UK, Sep. 2013.

2. Chavarria Reyes, E., Gutierrez-Estevez, D.M., and Akyildiz, I.F., "A Complete Femtocell Network Modeling and Development Platform," in Proc. of the IEEE Global Communications Conference (GLOBECOM), Anaheim, CA, Dec. 2012.

Patents

1. Akyildiz, I. F., Gutierrez-Estevez, D. M., and Chavarria Reyes, E. "Femtorelay Systems And Methods of Managing Same," US Patent Application No. 20120076027, Pub. Mar. 29, 2012.

REFERENCES

[1] Ericsson, "Ericsson Mobility Report," Technical Report, Jun. 2013.

[2] 3GPP, "Overview of 3GPP Release 8," Technical Report, Sep. 2013.

[3] International Telecommunication Union, "Requirements related to technical performance for IMT-Advanced radio interface(s)," Report ITU-R M.2134, 2008.

[4] 3GPP, "Requirements for further advancements for Evolved Universal Terrestrial Radio Access (E-UTRA) (LTE-Advanced)," Technical Report 36.913, Sep. 2012.

[5] I. F. Akyildiz, D. Gutierrez-Estevez, and E. Chavarria Reyes, "The Evolution to 4G Cellular Systems: LTE-Advanced," *Physical Communication (Elsevier) Journal*, vol. 3, no. 4, pp. 217–244, Dec. 2010.

[6] ArrayComm. Cooper's Law. [Online]. Available: http://www.arraycomm.com/technology/coopers-law

[7] J. R. Luening, "Femtocell Economics," in *GSMA Mobile World Conference*, Feb. 2009.

[8] J. Andrews, "Seven ways that HetNets are a cellular paradigm shift," *IEEE Communications Magazine*, vol. 51, no. 3, pp. 136–144, Mar. 2013.

[9] N. Bhushan, J. Li, D. Malladi, R. Gilmore, D. Brenner, A. Damnjanovic, R. Sukhavasi, C. Patel, and S. Geirhofer, "Network densification: the dominant theme for wireless evolution into 5g," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 82–89, Feb. 2014.

[10] T. E. Klein, "GreenTouch Consortium: Transforming ICT Networks for a Sustainable Future," Jan. 2013.

[11] Gartner, "Gartner Estimates ICT Industry accounts for 2 Percent of Global CO2 Emissions," Apr. 2007.

[12] The Boston Consulting Group, "GeSI SMARTer 2020: The Role of ICT in Driving a Sustainable Future," Global e-Sustainability Initiative, Tech. Rep., 2012.

[13] Alcatel-Lucent, "Sustainability Report A Business Imperative," Tech. Rep., 2012.

[14] Energy Aware Radio and neTwork tecHnologies. [Online]. Available: https://www.ict-earth.eu/

[15] Cognitive Radio and Cooperative Strategies for POWER saving in multi-standard wireless devices. [Online]. Available: http://www.ict-c2power.eu/

[16] Towards Real Energy-efficient Network Design. [Online]. Available: http://www.fp7-trend.eu/

[17] H. Claussen, L. T. W. Ho, and F. Pivit, "Effects of joint macrocell and residential picocell deployment on the network energy efficiency," in *Proc. of the IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Sep. 2008, pp. 1–6.

[18] U. Paul, A. Subramanian, M. Buddhikot, and S. Das, "Understanding traffic dynamics in cellular data networks," in *Proc. of the IEEE International Conference on Computer Communication (INFOCOM)*, Apr. 2011, pp. 882–890.

[19] R. Litjens and L. Jorguseski, "Potential of energy-oriented network optimisation: Switching off over-capacity in off-peak hours," in *Proc. of the IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Sep. 2010, pp. 1660–1664.

[20] B. Bougard, G. Lenoir, A. Dejonghe, L. Van der Perre, F. Catthoor, and W. Dehaene, "Smart mimo: An energy-aware adaptive mimo-ofdm radio link control for next-generation wireless local area networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2007, no. 3, p. 13, 2007.

[21] F. Cardoso, S. Petersson, M. Boldi, S. Mizuta, G. Dietl, R. Torrea-Duran, C. Desset, J. Leinonen, and L. Correia, "Energy efficient transmission techniques for LTE," *IEEE Communications Magazine*, vol. 51, no. 10, pp. 182–190, Oct. 2013.

[22] P. Rost and G. Fettweis, *Cooperative Cellular Wireless Networks*, 1st ed. Cambridge University Press, 2011.

[23] C. Bontu and E. Illidge, "DRX mechanism for power saving in LTE," *IEEE Communications Magazine*, vol. 47, no. 6, pp. 48–55, Jun. 2009.

[24] K. Chen and D. Peroulis, "Design of Adaptive Highly Efficient GaN Power Amplifier for Octave-Bandwidth Application and Dynamic Load Modulation," *IEEE Transactions on Microwave Theory and Techniques*, vol. 60, no. 6, pp. 1829–1839, 2012.

[25] H. K. Boyapati, R. Rajakumar, and S. Chakrabarti, "Quantifying the improvement in energy savings for LTE enodeb baseband subsystem with technology scaling and multi-core architectures," in *Proc. of the National Conference on Communication (NCC)*, Feb. 2012, pp. 1–5.

[26] M. Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo, "Optimal energy savings in cellular access networks," in *Proc. of the IEEE International Conference on Communications (ICC) Workshops*, Jun. 2009, pp. 1–5.

[27] F. Richter, A. Fehske, P. Marsch, and G. Fettweis, "Traffic demand and energy efficiency in heterogeneous cellular mobile radio networks," in *Proc. of the IEEE Vehicular Technology Conference (VTC)*, May 2010, pp. 1–6.

[28] Y. Qi, M. Imran, and R. Tafazolli, "Energy-aware adaptive sectorisation in LTE systems," in *Proc. of the IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Sep. 2011, pp. 2402–2406.

[29] G. Auer, V. Giannini, I. Godor, P. Skillermark, M. Olsson, M. Imran, D. Sabella, M. Gonzalez, C. Desset, and O. Blume, "Cellular Energy Efficiency Evaluation Framework," in *Proc. of the IEEE Vehicular Technology Conference (VTC)*, May 2011, pp. 1–6.

[30] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson, M. Imran, D. Sabella, M. Gonzalez, O. Blume, and A. Fehske, "How much energy is needed to run a wireless network?" *IEEE Wireless Communications Magazine*, vol. 18, no. 5, pp. 40–49, Oct. 2011.

[31] F. Han, Z. Safar, and K. Liu, "Energy-Efficient Base-Station Cooperative Operation with Guaranteed QoS," *IEEE Transactions on Communications*, vol. 61, no. 8, pp. 3505–3517, 2013.

[32] D. H. Ring, "Mobile telephony - wide area coverage - case 20564," Bell Telephone Laboratories Incorporated, Technical Memoranda, Dec. 1947.

[33] J. Andrews, F. Baccelli, and R. Ganti, "A Tractable Approach to Coverage and Rate in Cellular Networks," *IEEE Transactions on Communications*, vol. 59, no. 11, pp. 3122–3134, Nov. 2011.

[34] D. Cao, S. Zhou, and Z. Niu, "Optimal base station density for energy-efficient heterogeneous cellular networks," in *Proc. of the IEEE International Conference on Communications (ICC)*, Jun. 2012, pp. 4379–4383.

[35] K. Son, H. Kim, Y. Yi, and B. Krishnamachari, "Base Station Operation and User Association Mechanisms for Energy-Delay Tradeoffs in Green Cellular Networks," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 8, pp. 1525–1536, Sep. 2011.

[36] C. Desset, B. Debaillie, V. Giannini, A. Fehske, G. Auer, H. Holtkamp, W. Wajda, D. Sabella, F. Richter, M. Gonzalez, H. Klessig, I. Godor, M. Olsson, M. Imran, A. Ambrosy, and O. Blume, "Flexible power modeling of LTE base stations," in *Proc. of the IEEE Wireless Communications and Networking Conference (WCNC)*, Apr. 2012, pp. 2858–2862.

[37] E. Oh, B. Krishnamachari, X. Liu, and Z. Niu, "Toward dynamic energy-efficient operation of cellular network infrastructure," *IEEE Communications Magazine*, vol. 49, no. 6, pp. 56–61, Jun. 2011.

[38] H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam, "alpha-optimal user association and cell load balancing in wireless networks," in *Proc. of the IEEE International Conference on Computer Communication (INFOCOM)*, Mar. 2010, pp. 1–5.

[39] A. Conte, "Power consumption of base stations," Alcatel-Lucent Bell Labs France, TREND Plenary Meeting, Feb. 2012.

[40] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects," Technical Report 36.814, Mar. 2010.

[41] *Radio Regulations*, International Telecommunication Union Std., Nov. 2012. [Online]. Available: http://www.itu.int/pub/R-REG-RR/en

[42] K. Pedersen, F. Frederiksen, C. Rosa, H. Nguyen, L. Garcia, and Y. Wang, "Carrier aggregation for LTE-advanced: functionality and performance aspects," *IEEE Communications Magazine*, vol. 49, no. 6, pp. 89–95, Jun. 2011.

[43] Z. Shen, A. Papasakellariou, J. Montojo, D. Gerstenberger, and F. Xu, "Overview of 3GPP LTE-advanced carrier aggregation for 4G wireless communications," *IEEE Communications Magazine*, vol. 50, no. 2, pp. 122–130, Feb. 2012.

[44] 3GPP, "Service accessibility," Technical Specification 22.011, Dec. 2011.

[45] D. Bai, J. Lee, H. Nguyen, J. Singh, A. Gupta, Z. Pi, T. Kim, C. Lim, M.-G. Kim, and I. Kang, "LTE-advanced modem design: challenges and perspectives," *IEEE Communications Magazine*, vol. 50, no. 2, pp. 178–186, Feb. 2012.

[46] C. Park, L. Sundstro{}̈m, A. Wallen, and A. Khayrallah, "Carrier aggregation for LTE-advanced: design challenges of terminals," *IEEE Communications Magazine*, vol. 51, no. 12, pp. 76–84, Dec. 2013.

[47] A. Toskala, "Release 12 for c4 (cost, coverage, coordination with small cells and capacity)," in *TSG Ran Workshop on Rel-12 and onwards*. 3GPP, Jun. 2012.

[48] Huawei Technologies, "Views on Rel-12 and onwards for LTE and UMTS," in *TSG Ran Workshop on Rel-12*. 3GPP, Jun. 2012.

[49] NTT Docomo, "Requirements, Candidate Solution & Technology Roadmap for LTE Rel-12 Onward," in *TSG Ran Workshop on Rel-12*. 3GPP, Jun. 2012.

[50] C. Hoymann, D. Larsson, H. Koorapaty, and J.-F. Cheng, "A Lean Carrier for LTE," *IEEE Communications Magazine*, vol. 51, no. 2, pp. 74–80, Feb. 2013.

[51] Y. Wang, K. Pedersen, T. Sorensen, and P. Mogensen, "Carrier load balancing and packet scheduling for multi-carrier systems," *IEEE Transactions on Wireless Communications*, vol. 9, no. 5, pp. 1780–1789, May 2010.

[52] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. Andrews, "User Association for Load Balancing in Heterogeneous Cellular Networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 6, pp. 2706–2716, Jun. 2013.

[53] L. Venturino, A. Zappone, C. Risi, and S. Buzzi, "Energy-Efficient Scheduling and Power Allocation in Downlink OFDMA Networks with Base Station Coordination," *IEEE Transactions on Wireless Communications*, to be published, early Access.

[54] R. Hu and Y. Qian, "An energy efficient and spectrum efficient wireless heterogeneous network framework for 5g systems," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 94–101, May 2014.

[55] M. Ismail and W. Zhuang, "Network cooperation for energy saving in green radio communications," *IEEE Wireless Communications Magazine*, vol. 18, no. 5, pp. 76–81, Oct. 2011.

[56] O. Holland, A. Aghvami, T. Dodgson, and H. Bogucka, "Intra-operator dynamic spectrum management for energy efficiency," *IEEE Communications Magazine*, vol. 50, no. 9, pp. 178–184, Sep. 2012.

[57] B. Soret, H. Wang, K. Pedersen, and C. Rosa, "Multicell cooperation for LTE-advanced heterogeneous network scenarios," *IEEE Wireless Communications Magazine*, vol. 20, no. 1, pp. 27–34, Feb. 2013.

[58] C. D. T. Thai, P. Popovski, M. Kaneko, and E. de Carvalho, "Multi-Flow Scheduling for Coordinated Direct and Relayed Users in Cellular Systems," *IEEE Transactions on Communications*, vol. 61, no. 2, pp. 669–678, Feb. 2013.

[59] C.-L. I, C. Rowell, S. Han, Z. Xu, G. Li, and Z. Pan, "Toward green and soft: a 5g perspective," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 66–73, Feb. 2014.

[60] W. Liu, S. Han, and C. Yang, "Hybrid cooperative transmission in heterogeneous networks," in *Proc. of the IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Sep. 2012, pp. 921–925.

[61] K. Huang, S. Lu, and J. Guo, "An Optimized Cooperative Transmission Scheme for Interference Mitigation in Heterogeneous Downlink Network," in *Proc. of the IEEE Vehicular Technology Conference (VTC)*, Sep. 2012, pp. 1–5.

[62] B. Clerckx, Y. Kim, H. Lee, J. Cho, and J. Lee, "Coordinated multi-point transmission in heterogeneous networks: A distributed antenna system approach," in *Proc. of the IEEE International Midwest Symposium on Circuits and Systems (MWSCAS)*, Aug. 2011, pp. 1–4.

[63] J. Lee, Y. Kim, H. Lee, B. L. Ng, D. Mazzarese, J. Liu, W. Xiao, and Y. Zhou, "Coordinated multipoint transmission and reception in LTE-advanced systems," *IEEE Communications Magazine*, vol. 50, no. 11, pp. 44–50, Nov. 2012.

[64] Q. Li, R. Hu, Y. Qian, and G. Wu, "Cooperative communications for wireless networks: techniques and applications in LTE-advanced systems," *IEEE Wireless Communications Magazine*, vol. 19, no. 2, Apr. 2012.

[65] P. Marsch and G. Fettweis, "A decentralized optimization approach to backhaul-constrained distributed antenna systems," in *Proc. of the IST Mobile and Wireless Communications Summit (ISTMWC)*, Jul. 2007, pp. 1–5.

[66] ——, "A framework for optimizing the uplink performance of distributed antenna systems under a constrained backhaul," in *Proc. of the IEEE International Conference on Communications (ICC)*, Jun. 2007, pp. 975–979.

[67] 3GPP, "Enhanced CoMP for LTE," Samsung, Work Task RP-121803, Dec. 2012.

[68] ——, "Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Frequency (RF) system scenarios," Technical Report 36.942, Sep. 2012.

[69] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, March 2004.

[70] 3GPP, "Medium Access Control (MAC) protocol specification," Technical Specification 36.321, Mar. 2013.

[71] L. Zhou, H. Xu, H. Tian, Y. Gao, L. Du, and L. Chen, "Performance analysis of power saving mechanism with adjustable drx cycles in 3GPP LTE," in *Proc. of the IEEE Vehicular Technology Conference (VTC)*, 2008, pp. 1–5.

[72] S. Jin and D. Qiao, "Numerical analysis of the power saving in 3GPP LTE advanced wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 61, no. 4, pp. 1779–1785, 2012.

[73] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); LTE physical layer; General description," Technical Specification 36.201, Mar. 2011.

[74] ——, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures," Technical Specification 36.213, Mar. 2014.

[75] Y. F. Zhang, S. T. Gao, H. Tian, and B. Huang, "Delay analysis of DRX in LTE-advanced considering carrier aggregation," *The Journal of China Universities of Posts and Telecommunications*, vol. 18, no. 6, pp. 1 – 7, Dec. 2011.

[76] K. Zhou, N. Nikaein, and T. Spyropoulos, "Lte/lte-a discontinuous reception modeling for machine type communications," *IEEE Wireless Communications Letters*, vol. 2, no. 1, pp. 102–105, February 2013.

[77] T.-H. Lee, C.-J. Tsai, and T.-H. Wu, "Quality of service support under drx mechanism in LTE advanced wireless networks," in *Proc. of the IEEE Vehicular Technology Conference (VTC)*, 2013, pp. 1–5.

[78] H.-C. Wang, C.-C. Tseng, G.-Y. Chen, F.-C. Kuo, and K.-C. Ting, "Accurate analysis of delay and power consumption of LTE drx mechanism with a combination of short and long cycles," in *Proc. International Symposium on Wireless Personal Multimedia Communications (WPMC),*, 2012, pp. 384–388.

[79] S. Jha, A. Koc, R. Vannithamby, and M. Torlak, "Adaptive drx configuration to optimize device power saving and latency of mobile applications over LTE advanced network," in *Proc. of the IEEE International Conference on Communications (ICC)*, 2013, pp. 6210–6214.

[80] A. Koc, S. Jha, R. Vannithamby, and M. Torlak, "Device power saving and latency optimization in LTE-A networks through drx configuration," *IEEE Transactions on Wireless Communications*, 2014.

[81] C. Zhong, T. Yang, L. Zhang, and J. Wang, "A New Discontinuous Reception (DRX) Scheme for LTE-Advanced Carrier Aggregation Systems with Multiple Services," in *Proc. of the IEEE Vehicular Technology Conference (VTC)*, Sep. 2011, pp. 1–5.

[82] H. Takagi, *Queueing Analysis: A Foundation of Performance Evaluation*. Elsevier Science Publishers B.V., 1993, vol. 3.

[83] 3GPP, "Radio Link Control (RLC) protocol specification," Technical Specification 36.322, Sep. 2010.

[84] ——, "Packet Data Convergence Protocol (PDCP) specification," Technical Specification 36.323, Mar. 2013.

[85] ——, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description," Technical Specification 36.300, Dec. 2013.

[86] ——, "Radio Resource Control (RRC); Protocol specification," Technical Specification 36.331, Jun. 2014.

[87] ——, "Non-Access-Stratum (NAS) protocol for Evolved Packet System (EPS); Stage 3," Technical Specification 24.301, Mar. 2013.

[88] T. Zahir, K. Arshad, A. Nakata, and K. Moessner, "Interference management in femtocells," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 1, pp. 293–311, 2013.

[89] O. Tipmongkolsilp, S. Zaghloul, and A. Jukan, "The evolution of cellular backhaul technologies: Current issues and future trends," *IEEE Communications Surveys & Tutorials*, vol. 13, no. 1, pp. 97–113, 2011.

[90] I. F. Akyildiz, D. M. Gutierrez-Estevez, and E. Chavarria Reyes, "Method and Apparatus for Femto-Relays," U.S. Patent Application 20 120 076 027, Sep. 27, 2011.

[91] I. F. Akyildiz, E. Chavarria Reyes, D. M. Gutierrez-Estevez, Balakrishnan, and J. R. R., Krier, "Enabling Next Generation Small Cells through Femtorelays," in *Physical Communication (Elsevier) Journal*, Apr. 2013.

[92] D. Zhou and W. Song, "Interference-controlled load sharing with femtocell relay for macrocells in cellular networks," in *Proc. of the IEEE Global Telecommunications Conference (GLOBECOM)*, Dec. 2011, pp. 1–5.

[93] P. Jacob and A. Madhukumar, "Interference reduction through femto-relays," *IET Communications*, vol. 6, no. 14, pp. 2208–2217, 2012.

[94] ——, "Femto-relays: A power efficient coverage extension mechanism for femtocells," in *Proc. of the IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Sep. 2011, pp. 975–979.

[95] T. Elkourdi and O. Simeone, "Femtocell as a Relay: An Outage Analysis," *IEEE Transactions on Wireless Communications*, vol. 10, no. 12, pp. 4204–4213, Dec. 2011.

[96] A. Rath, S. Hua, and S. Panwar, "FemtoHaul: Using Femtocells with Relays to Increase Macrocell Backhaul Bandwidth," in *Proc. of the IEEE International Conference on Computer Communication (INFOCOM) Workshops*, Mar. 2010, pp. 1–5.

[97] S. Samarakoon, M. Bennis, W. Saad, and M. Latva-aho, "Enabling relaying over heterogeneous backhauls in the uplink of femtocell networks," in *Proc. of the International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt)*, May 2012, pp. 75–80.

[98] Y.-S. Chen, C.-C. Li, and W.-L. Chiang, "A femtocell-assisted data forwarding protocol in relay enhanced LTE networks," in *Proc. of the International Conference on Parallel Processing Workshops (ICPPW)*, Sep. 2011, pp. 127–136.

[99] A. Gamage, N. Rajatheva, and M. Codreanu, "Resource allocation for OFDMA-based relay assisted two-tier femtocell networks," in *Proc. of the International Symposium on Wireless Communication Systems (ISWCS)*, Nov. 2011, pp. 834–838.

[100] N. Nomikos, P. Makris, D. Skoutas, D. Vouyioukas, and C. Skianis, "A cooperation framework for LTE femtocells' efficient integration in cellular infrastructures based on femto relay concept," in *Proc. of the IEEE International Workshop on Computer Aided Modeling and Design of Commununication Links and Networks (CAMAD)*, Sep. 2012, pp. 318–322.

[101] D. Knisely, T. Yoshizawa, and F. Favichia, "Standardization of femtocells in 3GPP," *IEEE Communications Magazine*, vol. 47, no. 9, pp. 68–75, Sep. 2009.

[102] D. Knisely and F. Favichia, "Standardization of femtocells in 3GPP2," *IEEE Communications Magazine*, vol. 47, no. 9, pp. 76–82, Sep. 2009.

[103] 3GPP, "UTRAN Iu interface: General aspects and principles," Technical Specification 25.410, Sep. 2012.

[104] ——, "UTRAN architecture for 3G Home Node B (HNB)," Technical Specification 25.467, Dec. 2011.

[105] ——, "UTRAN Iuh Interface RANAP User Adaption (RUA) signalling," Technical Specification 25.468, Dec. 2012.

[106] ——, "UTRAN Iuh interface Home Node B (HNB) Application Part (HNBAP) signalling," Technical Specification 25.469, Jun. 2012.

[107] *Information technology - Abstract syntax Notation One (ASN.1): Specification of basic notation*, Telecommunication Standardization Sector ITU-T Recommendation X.6880, November 2008.

[108] *Information technology - ASN.1 encoding rules: Specification of Basic Encoding Rules (BER), Canonical Encoding Rules (CER) and Distinguished Encoding Rules (DER)*, Telecommunication Standardization Sector ITU-T Recommendation X.690, November 2008.

[109] T. Szigeti and C. Hattingh. (2004, Dec.) Quality of Service Design Overview. Cisco. [Online]. Available: http://www.ciscopress.com/articles/article.asp?p=357102

[110] J. Sydir and R. Taori, "An evolved cellular system architecture incorporating relay stations," *IEEE Communications Magazine*, vol. 47, no. 6, pp. 115–121, Jun. 2009.

[111] C. Raman, G. Foschini, R. Valenzuela, R. Yates, and N. B. Mandayam, "Half-Duplex Relaying in Downlink Cellular Systems," *IEEE Transactions on Wireless Communications*, vol. 10, no. 5, pp. 1396–1404, May 2011.

[112] Krishnan, N. and Yates, R.D. and Mandayam, Narayan B. and Panchal, J.S., "Bandwidth Sharing for Relaying in Cellular Systems," *IEEE Transactions on Wireless Communications*, vol. 11, no. 1, pp. 117–129, Jan. 2012.

[113] P. Xia, V. Chandrasekhar, and J. Andrews, "Open vs. closed access femtocells in the uplink," *IEEE Transactions on Wireless Communications*, vol. 9, no. 12, pp. 3798–3809, Dec. 2010.

[114] H.-S. Jo, P. Xia, and J. Andrews, "Downlink Femtocell Networks: Open or Closed?" in *Proc. of the IEEE International Conference on Communications (ICC)*, Jun. 2011, pp. 1–5.

[115] S.-Y. Yun, Y. Yi, D.-H. Cho, and J. Mo, "The Economic Effects of Sharing Femtocells," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 3, pp. 595–606, Apr. 2012.

[116] R. Singoria, T. Oliveira, and D. Agrawal, "Reducing Unnecessary Handovers: Call Admission Control Mechanism between WiMAX and Femtocells," in *Proc. of the IEEE Global Telecommunications Conference (GLOBECOM)*, Dec. 2011, pp. 1–5.

[117] T. Guo, A. ul Quddus, and R. Tafazolli, "Seamless Handover for LTE Macro-Femto Networks Based on Reactive Data Bicasting," *IEEE Communications Letters*, vol. 16, no. 11, pp. 1788–1791, Nov. 2012.

[118] T. Guo, A. ul Quddus, N. Wang, and R. Tafazolli, "Local Mobility Management for Networked Femtocells Based on X2 Traffic Forwarding," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 1, pp. 326–340, Jan. 2013.

[119] T. Riihonen, S. Werner, and R. Wichman, "Comparison of Full-Duplex and Half-Duplex Modes with a Fixed Amplify-and-Forward Relay," in *Proc. of the IEEE Wireless Communications and Networking Conference (WCNC)*, Apr. 2009, pp. 1–5.

[120] V. Cadambe and S. Jafar, "Degrees of Freedom of Wireless Networks With Relays, Feedback, Cooperation, and Full Duplex Operation," *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2334–2344, May 2009.

[121] Y. Y. Kang and J. H. Cho, "Capacity of MIMO wireless channel with full-duplex amplify-and-forward relay," in *Proc. of the IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Sep. 2009, pp. 117–121.

[122] Q. Li, K. Li, and K. Teh, "Achieving Optimal Diversity-Multiplexing Tradeoff for Full-Duplex MIMO Multihop Relay Networks," *IEEE Transactions on Information Theory*, vol. 57, no. 1, pp. 303–316, Jan. 2011.

[123] D. Ng and R. Schober, "Dynamic Resource Allocation in OFDMA Systems with Full-Duplex and Hybrid Relaying," in *Proc. of the IEEE International Conference on Communications (ICC)*, Jun. 2011, pp. 1–6.

[124] T. Riihonen, S. Werner, and R. Wichman, "Hybrid Full-Duplex/Half-Duplex Relaying with Transmit Power Adaptation," *IEEE Transactions on Wireless Communications*, vol. 10, no. 9, pp. 3074–3085, Sep. 2011.

[125] A. Host-Madsen and J. Zhang, "Capacity bounds and power allocation for wireless relay channels," *IEEE Transactions on Information Theory*, vol. 51, no. 6, pp. 2020–2040, Jun. 2005.

[126] H. Ju, E. Oh, and D. Hong, "Improving efficiency of resource usage in two-hop full duplex relay systems based on resource sharing and interference cancellation," *IEEE Transactions on Wireless Communications*, vol. 8, no. 8, pp. 3933–3938, Aug. 2009.

[127] P. Lioliou, M. Viberg, M. Coldrey, and F. Athley, "Self-interference suppression in full-duplex MIMO relays," in *Proc. of the Conference on Signals, Systems and Computers (ASILOMAR)*, Nov. 2010, pp. 658–662.

[128] T. Riihonen, S. Werner, and R. Wichman, "Mitigation of Loopback Self-Interference in Full-Duplex MIMO Relays," *IEEE Transactions on Signal Processing*, vol. 59, no. 12, pp. 5983–5993, Dec. 2011.

[129] M. Duarte, C. Dick, and A. Sabharwal, "Experiment-Driven Characterization of Full-Duplex Wireless Systems," *IEEE Transactions on Wireless Communications*, vol. 11, no. 12, pp. 4296–4307, Dec. 2012.

[130] W. Schacherbauer, A. Springer, T. Ostertag, C. Ruppel, and R. Weigel, "A flexible multiband frontend for software radios using high IF and active interference cancellation," in *Proc. of the IEEE MTT-S International Microwave Symposium*, vol. 2, May 2001, pp. 1085–1088.

[131] M. Duarte and A. Sabharwal, "Full-duplex wireless communications using off-the-shelf radios: Feasibility and first results," in *Proc. of the Conference on Signals, Systems and Computers (ASILOMAR)*, Nov. 2010, pp. 1558–1562.

[132] J. I. Choi, M. Jain, K. Srinivasan, P. Levis, and S. Katti, "Achieving Single Channel, Full Duplex Wireless Communication," in *Proc. of the International Conference on Mobile Computing and Networking (MOBICOM)*. New York, NY, USA: ACM, 2010, pp. 1–12.

[133] Intersil. Qhx220 active isolation enhancer and interference canceller. Intersil. [Online]. Available: http://www.intersil.com/en/products/other-analog/noise-canceller/isolation-enhancer-noise-cancellation/QHX220.html?pn=QHX220

[134] B. Widrow, J. Glover, J.R., J. McCool, J. Kaunitz, C. Williams, R. Hearn, J. Zeidler, J. Eugene Dong, and R. Goodlin, "Adaptive noise cancelling: Principles and applications," *Proceedings of the IEEE*, vol. 63, no. 12, pp. 1692–1716, Dec. 1975.

[135] J. Yuan, J. Shi, B. Tang, and H. Chen, "An Adaptive Feedback Interference Cancelling Algorithm Based on Independent Component Analysis for Wireless Repeaters," in *Proc. of the International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM)*, Sep. 2010, pp. 1–4.

[136] A. Sahai, G. Patel, and A. Sabharwal, "Pushing the limits of Full-duplex: Design and Real-time Implementation," Rice University, Technical Report TREE1104, Jun. 2011.

[137] 3GPP, "Evolved Universal Terrestrial Radio Access Network (E-UTRAN); X2 general aspects and principles," Technical Specification 36.420, Sep. 2012.

[138] ——, "Evolved Universal Terrestrial Radio Access Network (E-UTRAN); S1 general aspects and principles," Technical Specification 36.410, Sep. 2012.

VITA

Elías Chavarría Reyes received the Engineering Degree in Electronics and Communication from Universidad de Panamá, Ciudad de Panamá, Panamá, in 2007. From 2007 to 2008, he worked in Centauri Technologies Corporation as Junior Analyst. During this period, he was in charge of the analysis and implementation of information security systems. In May 2010, he received his Master of Science in Electrical and Computer Engineering from the Georgia Institute of Technology, Atlanta. From June to August 2013, he was an intern in Alcatel-Lucent Bell Labs focusing on the development of cellular network data acquisition and processing tools. Since August 2009, Elías has been a Ph.D. student with the Broadband Wireless Networking Lab at the Georgia Institute of Technology, under the supervision of Professor Ian F. Akyildiz, with a fellowship of "SENACYT" for the years 2008-2013. Elías was the recipient of the Oscar P. Cleaver Award for outstanding graduate students in the School of Electrical and Computer Engineering, at the Georgia Institute of Technology in 2009. His research interests include next generation cellular systems, particularly the area of heterogeneous wireless systems and their energy consumption. Elías is a student member of IEEE.